UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Tarifação de seguros multirriscos habitação

Victor Ricardo Pestana de Abreu

Mestrado em Matemática Aplicada à Economia e Gestão

Versão Pública

Trabalho de Projeto orientado por:
Professora Doutora Maria Teresa Alpuim

2020

Agradecimentos

Estes agradecimentos destinam-se a todos aqueles que me ajudaram a trilhar um caminho de sucesso.

Em primeiro lugar, gostaria de agradecer à minha família - em particular, aos meus pais, por através dos seus esforços terem tornado possível a minha presença neste Mestrado, e à minha irmã, pelo suporte que me deu nesta nova fase da minha vida.

Gostaria também de agradecer à Professora Doutora Teresa Alpuim, pelo apoio que me deu na elaboração deste trabalho, estando sempre disponível para fornecer informações e esclarecer as dúvidas que tive ao longo deste projeto. Adicionalmente, agradeço a todos os professores que, ao terem leccionado disciplinas por mim frequentadas, contribuíram para o enriquecimento dos meus conhecimentos.

Tendo sido este um trabalho desenvolvido no contexto de um estágio, não poderia deixar de agradecer o contributo de todos os membros da Aegon Santander Portugal com os quais tive o prazer de interagir, sendo este agradecimento destinado sobretudo à Cláudia Conceição, quem sempre esteve disponível para me guiar neste trajeto. Gostaria ainda de agradecer ao Tiago Venâncio e à Susana Correia, pela confiança que depositaram em mim desde o início do estágio.

Por último, aproveito ainda para agradecer também a todas as restantes entidades ainda não mencionadas mas que, de uma forma ou de outra, tornaram possível o meu sucesso - como os meus amigos e colegas de curso e a própria Universidade de Lisboa, de uma forma geral.

Resumo

O presente trabalho, realizado no âmbito de um estágio profissional na Aegon Santander Portugal (ASP), visa resultar na criação de uma estrutura tarifária para o produto de seguros multirriscos habitação, estrutura essa obtida através do recurso a conjuntos de dados adequados e a modelos lineares generalizados, os quais representam opções viáveis para incorporar informação presente em covariáveis na modelação de riscos, ajudando a eliminar em grande parte qualquer possibilidade de seleção adversa.

Neste sentido, será exemplificada a construção da tarifa pretendida, não sem antes serem introduzidos conceitos básicos sobre Seguros e Atuariado, sobre a ASP e sobre os seguros multirriscos habitação no geral. Apesar da natureza prática deste trabalho, decidiu-se também apresentar de forma breve a teoria dos modelos lineares generalizados, antes do uso dos mesmos na modelação tarifária em Atuariado. Por último, mas não menos importante, é dada alguma atenção à implementação dos modelos criados, com o intuito de avaliar a suficiência dos prémios cobrados.

Palavras-chave: Atuariado, seguros multirriscos habitação, tarifação, modelos lineares generalizados.

Abstract

The present paper, carried out as part of a professional internship at Aegon Santander Portugal (ASP), has as its purpose the creation of a tariff structure for the multirisk home insurance product. This exact structure was obtained through the use of appropriate data sets and generalized linear models, which represent viable options to incorporate the information present in covariates in risk modeling, effectively eliminating, in large proportion, any possibility of adverse selection.

The construction of the intended tariff will be explained, not without first introducing some basic concepts about Insurance and Actuarial Science, ASP and multirisk home insurance. Despite the practical nature of this paper, it is important to mention that the theory of generalized linear models will also be presented in a short but comprehensive way, before its use in tariff building. Last but not least, the implementation of the created models will be addressed with some emphasis. The goal is to evaluate if the premiums charged are adequate and reasonable.

Keywords: Actuarial science, multirisk home insurance, pricing, generalized linear models.

Índice

Lista de figuras

Lista de tabelas

Introdução

Motivação, enquadramento, estrutura e objetivos deste trabalho

Este Trabalho de Projeto foi realizado no âmbito do Mestrado em Matemática Aplicada à Economia e Gestão da Faculdade de Ciências da Universidade de Lisboa, e no contexto de um estágio profissional realizado na companhia de seguros Aegon Santander Portugal (doravante, ASP), visando a aplicação de técnicas quantitativas e de conhecimentos científicos adquiridos em contexto académico na resolução de problemas em contexto profissional.

Mais concretamente, este trabalho é dedicado à temática da tarifação de seguros, neste caso de um seguro do ramo patrimonial – o seguro de multirriscos habitação – comercializado pela ASP sob a designação Proteção Lar. Para tal, considerou-se a aplicação de modelos estatísticos que permitissem distinguir apólices com base nos riscos que estes representam para a ASP, permitindo assim estimar os prémios (puros) a aplicar a cada apólice, de acordo com as suas caraterísticas, surgindo assim a ideia da aplicação de Modelos Lineares Generalizados. No entanto, esta não foi uma sugestão aplicada de forma "cega", sendo antes o resultado da identificação das necessidades da ASP aquando do início do estágio.

Porém, e como em qualquer projeto que envolva análise de dados, e para além dos conteúdos probabilísticos e estatísticos relevantes, dos conceitos e raciocínios matemáticos que lhes servem de base e ainda da implementação computacional da dita análise, importa conhecer o contexto da aplicação. Esta preocupação é especialmente válida na área dos seguros, com especificidades próprias tanto ao nível dos termos empregues, como ao nível do próprio negócio. Por isso, e para reconhecer esta área de aplicação, iremos falar em modelos atuariais (sem prejuízo destes modelos serem também estatísticos). Também por isso, e antes de quaisquer secções dedicadas a conteúdos estatísticos, é feita (nos capítulos iniciais) uma introdução breve à atividade seguradora, à ASP e aos seguros multirriscos habitação, com o intuito de enquadrar este trabalho, de natureza quantitativa, com o contexto no qual o mesmo se insere e do qual é inseparável.

Só depois desta introdução é abordada a temática dos Modelos Lineares Generalizados. Como é da opinião do autor que os problemas que surgem no contexto da Matemática Aplicada merecem uma resolução tão rigorosa, eficiente e interpretável quanto possível, e que tal resolução exige um profundo conhecimento dos conceitos a aplicar, é apresentada primeiro a teoria da área relevante para o contexto de aplicação, sendo depois explicada a aplicação destes conceitos em si, e no contexto do estágio profissional realizado.

Tendo em mente que o objetivo de qualquer projeto de modelação matemática é o de utilização de modelos pertinentes, e não a mera criação dos mesmos, torna-se necessário interpretar o significado de tais modelos. Só depois deste estudo será possível extrair conclusões deste Trabalho de Projeto e apresentar a bibliografia e os anexos relevantes.

Os objetivos primários deste trabalho passam então:

- Por criar um modelo tarifário que seja – até onde permitido pelos dados – interpretável e útil para efeitos de previsão;

- Por aplicar o modelo tarifário construído para estimação de prémios puros, melhorando o *modus operandi* atualmente em vigor, trazendo assim valor para a ASP.

Adicionalmente, os objetivos secundários deste trabalho são, para o autor, os seguintes:

- Desenvolver conhecimentos relevantes e diretamente aplicáveis no exercício regular de funções atuariais na atividade da Aegon Santander Portugal;
- Ampliar os conhecimentos matemáticos e estatísticos oriundos da formação académica do autor e das suas iniciativas de aprendizagem;
- Despertar uma maior atenção para questões de eficiência e complexidade computacional, tendo em mente possíveis análises futuras envolvendo grandes volumes de dados;
- Refinar e aplicar metodologias integradas de análise de dados;
- Desenvolver capacidades de raciocínio lógico e de espírito crítico;
- Exercitar aptidões de escrita de documentos em linguagem fluida e correta, e num formato adequado;
- Desenvolver um trabalho que sirva de referência para estudantes de áreas quantitativas e que possa também ser útil para o público em geral.

Por último, e agora na perspetiva da ASP, há ainda três objetivos terciários a salientar, os quais passam por:

- Desenvolver capacidades de resolução de problemas com informação limitada e/ou com restrições de tempo;
- Sugerir ações de negócio e desenvolver conhecimentos económicos e empresariais;
- Levantar questões e sugestões sobre a qualidade e pertinência dos dados e do seu armazenamento.

Nota: por motivos de confidencialidade e de proteção de dados, foi necessário anonimizar alguns conteúdos desta tese, nomeadamente nomes de variáveis (e suas categorias), montantes monetários e outros aspetos referentes à atividade interna da ASP.

Seguros e Atuariado: conceitos básicos

Funções e benefícios associados a seguros. Falhas de mercado - o risco moral e a seleção adversa

Todas as pessoas necessitam e têm necessitado, desde tempos imemoriais, de segurança económica (na forma de abrigo, alimentos, roupa, cuidados médicos, etc.), para si e para os seus bens/posses (ou para os da sua empresa, por exemplo), sobretudo para fazer face a fenómenos aleatórios e fora do seu controlo que levem à perda de tal segurança como, por exemplo, catástrofes naturais. Esta possibilidade nefasta representa, perante a existência de probabilidades associadas, uma manifestação de risco económico, ou simplesmente risco, o qual pode, no limite, ameaçar diretamente a sua vida e/ou a sua atividade económica.

É este receio de que algo mau aconteça, e de que daí advenham consequências negativas, que motiva muitas pessoas e organizações a adquirir seguros, os quais são acordos escritos que envolvem pelo menos duas entidades - o tomador do seguro e a companhia de seguros, ou seguradora - e nos quais o tomador aceita pagar à seguradora pelo menos um montante determinístico previamente estipulado, chamado de prémio. Em troca, e dependendo do tipo de seguro adquirido (e das suas caraterísticas), perante a ocorrência de eventos aleatórios com consequências financeiras desfavoráveis, o beneficiário deste seguro pode receber ou solicitar pagamentos (dentro dos limites contratados), entre outras prestações acordadas,

- Pelo menos uma indemnização, de natureza monetária, no caso de ocorrência de fenómenos dos quais resultem prejuízos materiais;
- A prestação de serviços, os quais servem também para fazer face a estes prejuízos e/ou a outras perdas;
- Pelo menos um montante no formato de capital ou de renda, em caso de acontecimento respeitante à pessoa humana;

Tendo tais eventos desfavoráveis natureza aleatória, os impactos por eles causados serão também aleatórios.

De forma mais simples, seguros servem para indivíduos e organizações adquirirem proteção para perdas causadas por possíveis infortúnios, a troco do pagamento de um preço - o prémio. Para este prémio poder ser considerado justo terá, como veremos ao longo deste documento, de ser calculado em função do risco que cada indivíduo ou entidade traz à seguradora. Porém, podemos desde já dizer que o prémio do seguro deve ser baixo ao ponto de possibilitar esta troca, isto é, de ser percecionado como sendo vantajoso pelo consumidor.

Através da aquisição de uma apólice de seguro, cada tomador troca uma perda possivelmente avultada (embora não garantida) pelo pagamento de um prémio significativamente menor (embora garantido) e, nesse sentido, um seguro pode ser apelativo, na medida em que permite ao tomador cuidar da sua situação financeira futura (ou da de alguém que lhe é próximo) não através da procura de rendimentos, mas através da redução do potencial para a ocorrência de perdas ruinosas.

Os montantes a pagar pela seguradora em caso de materialização de riscos podem ser valores previamente determinados ou valores calculados em função da perda ocorrida, sendo tal perda reembolsada

no todo ou em parte. Tais perdas serão, se realizadas, mais facilmente suportadas pela seguradora, através da agregação dos riscos associados a muitos indivíduos e empresas, pois estes montantes serão extraídos do total de prémios pagos pelos detentores das apólices e, desse modo, o impacto financeiro de um evento que poderia ser desastroso para um tomador é dividido por um grupo maior, ou seja, "todos acabam por pagar pelo infortúnio de alguns". Por isso, seguros, enquanto mecanismos de transferência de risco, são baseados na partilha dos mesmos, assentando por sua vez tal partilha neste princípio de mutualidade, de caráter essencial.

O agrupamento de riscos e perdas não significa necessariamente que cada membro de um grupo (*pool*) de risco exerça a mesma contribuição. Na prática, estes grupos costumam conter membros com maior nível de risco (os quais pagam no geral maiores prémios, pois é mais provável que façam com que a seguradora lhes tenha de prestar benefícios), e outros com menor nível de risco (os quais tendem a pagar menos, por motivos similares). Estas medidas incentivam os membros com menor risco a permanecer no grupo.

A mutualidade é então um princípio fundamental da atividade seguradora de natureza privada e comercial, definido por David Wilkie em 1997, o qual afirma que perdas devem ser divididas entre os tomadores de seguro, os quais pagam prémios de acordo com o risco que trazem consigo no momento de celebração do contrato de seguro. Tal conceito não deve, contudo, ser confundido com o paradigma de solidariedade, também definido por Wilkie, no qual se defende a partilha de perdas através do pagamento de contribuições de acordo com os rendimentos de cada indivíduo; é esta a natureza de sistemas públicos de segurança social, de natureza universal e obrigatória.

Por sua vez, a apólice de seguro é a evidência documental do contrato de seguro celebrado entre a companhia de seguros e o tomador do seguro, a qual indica as condições do contrato acordadas entre as partes (gerais, especiais, se as houver, e particulares). Tal apólice, enquanto registo escrito da celebração do contrato de seguro, é obrigatória por lei, mais concretamente pelo Artigo 32º. da Secção V do Decreto-Lei nº. 72/2008.

O prémio é outro elemento essencial para o contrato de seguro, sendo definido como o valor total (incluindo taxas e impostos) da contribuição que o tomador do seguro sujeito a riscos deve pagar (nos prazos indicados) ao segurador, para poder beneficiar de proteção face a certas ocorrências através da aquisição de um seguro.

Por sua vez, e de forma lata, um risco é uma condição na qual mais de um resultado é possível, sendo um ou mais destes resultados negativos. Mais concretamente, por risco entende-se a possibilidade de ocorrência futura de um acontecimento desfavorável do qual resultam perdas, danos (como morte) e/ou necessidades económicas, para os quais se pode pretender cobertura. De notar que, não existindo uma definição universalmente aceite de risco, este conceito é frequentemente confundido com o de incerteza. Assim, será adotada neste documento a definição transmitida por Frank H. Knight no seu livro de 1921, *Risk, Uncertainty, and Profit*, do qual se destacam as seguintes frases:

- "*Risk is present when future events occur with measurable probability*";
- "*Uncertainty is present when the likelihood of future events is indefinite or incalculable*".

Ou seja, é adotada aqui a ideia comumente utilizada em Finanças/Economia/Gestão de que risco consiste em *conhecer o desconhecido*, através de probabilidades (as quais quantificam a aleatoriedade), ao

contrário de incerteza, a qual implica *desconhecer o desconhecido* e na qual é, portanto, impossível quantificar a aleatoriedade.

De sublinhar que o risco é um elemento essencial do contrato de seguro, no sentido em que sem risco não existe (contrato de) seguro; só é seguro o contrato no qual se faz referência a um risco, tendo este caráter essencial (como descrito no Artigo 44º. do Regime Jurídico do Contrato de Seguro).

O risco é então o elemento que motiva a existência do contrato de seguro. Este risco é transferido, através do contrato de seguro, dos indivíduos (inicialmente expostos a perdas) para a seguradora, sendo esta transferência justificada pelo facto de tal seguradora reduzir o seu risco ao comercializar seguros a um número suficientemente elevado de indivíduos, de maneira a conseguir prever de forma mais precisa/exata o montante das suas perdas totais.

Muitas atividades que damos como adquiridas envolvem riscos e poderiam não ser realizadas sem a existência de seguros como, por exemplo, de acidentes de trabalho. Adicionalmente, sem seguros, as pessoas teriam mais dificuldade em criar o seu próprio negócio ou adquirir a sua própria habitação, devido aos potenciais custos aos quais estariam totalmente expostos. Neste sentido, os seguros cumprem uma função social importante.

Seguros funcionam melhor quando os riscos cobertos são partilhados por grupos de maiores dimensões. No entanto, a partilha destes riscos não deve ser feita de qualquer forma, uma vez que seguros também estão associados a falhas de mercado, como a seleção adversa e o risco moral.

Seleção adversa (também denominada de anti seleção) ocorre quando indivíduos com maiores exposições ao risco têm (em comparação com indivíduos associados a menores riscos) maior probabilidade de adquirir proteção para os mesmos através de seguros, e tende a ocorrer quando os tomadores de seguro, segurados e/ou beneficiários conhecem melhor os seus próprios riscos do que a própria seguradora.

Dá-se a existência de seleção adversa em contextos onde todos os tomadores de seguros pagam o mesmo prémio, pois aí os tomadores de menor (maior) risco estão a pagar mais (menos) do que o valor actuarialmente justo do prémio (respetivamente). Isto conduz à saída dos indivíduos de menor risco do grupo, e também à entrada de indivíduos de maior risco no mesmo, o que pode levar (no limite) ao colapso da seguradora.

Para evitar a ocorrência de seleção adversa, as seguradoras procuram efetuar uma avaliação de riscos que seja o mais precisa possível, cobrando um prémio ajustado ao verdadeiro nível de risco associado a cada segurado, assim distinguindo tais níveis entre segurados. Uma boa classificação de riscos deve então identificar fatores de risco relevantes e, com base nestes (os quais alimentam modelos quantitativos) estimar prémios que reflitam corretamente os riscos aos quais dizem respeito.

Uma tarifação eficiente será portanto baseada nos riscos, de forma a desencorajar seleção adversa, existindo no entanto dois obstáculos à realização da mesma, por parte das seguradoras:

- Podem existir restrições na disponibilidade/capacidade de obtenção de dados adequados, que permitam avaliar corretamente o risco de cada indivíduo;
- Podem não ser aprovados, por parte dos reguladores e demais entidades governamentais, certos aspetos na construção de modelos tarifários (como, por exemplo, a discriminação baseada em género), dependendo tais decisões das perceções que os mesmos têm de eficiência e de equidade.

Por sua vez, estamos perante um problema de risco moral quando a existência de um contrato entre duas partes provoca numa delas uma mudança de comportamento que fere ou prejudica o bem-estar da outra parte.

Na Atividade Seguradora, diz-se que ocorre risco moral se a existência de seguros motiva os segurados a mudar os seus comportamentos, reduzindo as suas precauções (as quais seriam exercidas na inexistência de seguro) e tornando-os menos prudentes, de um modo que torna a eventualidade do sinistro mais passível de ocorrer, motivando assim perdas quer por desleixo, quer de forma propositada.

O risco moral pode resultar em mais indemnizações do que a seguradora esperava, na sequência do seu processo de subscrição. Seguros são inviáveis nos casos mais extremos de risco moral, nos quais perdas são intencionais e, portanto, não-aleatórias. Se as seguradoras tiverem de assumir que, no geral, os indivíduos segurados terão comportamentos mais imprudentes, a consequência mais óbvia será o aumento de prémios, com vista à cobertura de maiores perdas esperadas.

Os conceitos de seleção adversa e de risco moral correspondem, como já visto, a falhas de mercado, no sentido em que as restrições no acesso a informações por parte das seguradoras pode levar a que os seguros oferecidos pelo mercado sejam não ótimos, levando a colapsos em casos extremos. Assim, é sem surpresas que nos apercebemos que as seguradoras se esforçam para obter acesso a cada vez mais dados sobre os riscos que os segurados representam, aplicando de seguida processos de subscrição (*underwriting*), precificação (*pricing*), e conceção da apólice (*policy-design*), os quais visam desincentivar estas falhas e, portanto, ajudar ao correto funcionamento de mercados.

Prémios, tarifas, riscos e sinistros. Especificidades do setor segurador

Como visto anteriormente, os prémios cobrados a tomadores de seguro deverão depender da magnitude do risco que estes representam. No entanto, nem todos os riscos serão precificados, pois nem todos os riscos são seguráveis. Para um dado risco ser segurável, este deve satisfazer (pelo menos em teoria) os seguintes critérios:

- Os riscos devem poder ser claramente definidos e financeiramente quantificados;
- Os riscos devem ser aleatórios e independentes;
- O segurado deve ter um interesse segurável;
- A seguradora deve poder calcular um prémio de risco que seja justo e viável ou, por outras palavras, que faça sentido economicamente;
- O risco deve ter probabilidade calculável, se possível com um número suficientemente elevado de unidades expostas ao risco (de preferência identicamente distribuídas);
- O risco de perdas catastróficas deve ser limitado ou inexistente.

Na realidade, são poucos os riscos que verificam de forma cumulativa e exata todas estas condições, mas quanto mais se afastarem destas, menos seguráveis se tornam.

As seguradoras procuram então segurar riscos puros - aqueles onde o indivíduo, na melhor das hipóteses, permanece na mesma situação de riqueza seja quais forem os eventos ocorridos - em detrimento de riscos especulativos - nos quais o indivíduo pode, pelo menos em certas circunstâncias, utilizar o seguro para obter lucros, aos quais não teria acesso na inexistência dos mesmos.

A seguradora pode também impor restrições aos riscos cobertos - por exemplo, especificando quais os riscos que decide cobrir, ou excluindo certos riscos. Iremos então assumir, ao longo deste documento, que os riscos com os quais nos deparamos são de facto seguráveis e do interesse da seguradora – e portanto precificáveis.

Às dificuldades inerentes à definição de preços, juntam-se as especificidades da atividade seguradora, a qual possui habitualmente um ciclo de produção invertido, no sentido em que os recebimentos (*cash-flows* estáveis, na forma de prémios) antecedem os pagamentos (*cash-flows* incertos, na forma de benefícios), isto é, são cobrados primeiro os prémios, para só depois se proceder à prestação de eventuais benefícios, sendo o montante e a própria ocorrência futura destes benefícios de natureza aleatória e, por vezes, conhecida apenas anos depois.

Assim sendo, é necessária a cobrança prévia de montantes que sejam suficientes para cobrir, com probabilidade elevada, os montantes que a seguradora terá de pagar a nível de sinistros no futuro (em adição a despesas da própria seguradora, por exemplo, de natureza operacional). Este é, portanto, um dos principais desafios da atividade seguradora - ao contrário de outros negócios mais tradicionais, nos quais "basta" vender os produtos/serviços acima dos custos que lhes são imputados, custos estes conhecidos com certeza (quase) total no momento do estabelecimento do preço de venda. Também por isso é necessário um maior grau de controlo financeiro das mesmas, o que leva à existência de autoridades de supervisão, as quais têm uma preocupação fundamental com clientes de companhias de seguros e, consequentemente, com a solvência financeira destas companhias.

Existem, no entanto, montantes de diversas naturezas que um tomador pode ter de pagar à seguradora como contrapartida das obrigações assumidas pela mesma; por isso, prémios compreendem cinco parcelas:

1. O prémio puro - é o prémio considerado estritamente necessário pela seguradora para satisfazer pedidos de indemnizações associados aos riscos que aceitou, sendo calculado através da condição de equilíbrio atuarial, a qual afirma que o valor atual esperado das quantias a pagar pela seguradora deve ser igual ao valor atual esperado das quantias a pagar pelo tomador do seguro;
2. A carga ou margem de segurança - a qual permite à seguradora criar uma reserva a utilizar em anos menos favoráveis e assim minimizar a probabilidade de ruína da mesma (uma vez que o valor efetivamente pago em sinistros pode ultrapassar o valor esperado dos mesmos);
3. As cargas de aquisição, gerência e cobrança - servem para fazer face aos encargos da seguradora e destinam-se, por exemplo, ao pagamento de comissões e à cobertura de custos de conceção do produto (isto é, de tratamento da proposta de seguro e do processo de subscrição) e de manutenção do mesmo, bem como ao pagamento de custos relacionados com a gestão de sinistros;
4. A carga de fracionamento - aplicável quando a seguradora facilita o pagamento dos prémios anuais (calculados assumindo que serão pagos no início de cada ano) em prestações mensais, trimestrais, quadrimestrais ou semestrais (ou seja, correspondentes a uma fração de um ano);
5. Os impostos e taxas legais - os quais[1] incidem sobre prémios de seguros.

Assim sendo, existem vários tipos de prémio na atividade seguradora, os quais são indicados de seguida:

[1] Como imposto do selo (IS), taxas parafiscais como contribuições ao INEM (Instituto Nacional de Emergência Médica), FAT (Fundo Acidentes de Trabalho), ANPC (Autoridade Nacional de Proteção Civil), FGA (Fundo de Garantia Automóvel), Serviço Nacional de Bombeiros e Proteção Civil (SNBPC) ou Certificado Responsabilidade Civil (CRC).

1. Prémio puro ou prémio de risco (ou prémio estatístico, ou ainda custo técnico), o qual, como já visto, reflete apenas os custos da cobertura do risco;
2. Prémio de inventário - resulta da adição de encargos administrativos/de gestão ao prémio puro;
3. Prémio comercial (ou prémio líquido) - resulta da adição dos encargos comerciais/de cobrança e dos encargos de aquisição ao prémio de inventário;
4. Prémio bruto - resulta da adição de encargos de fracionamento (caso estes existam) ao prémio comercial;
5. Prémio total – é o prémio cobrado pela seguradora ao tomador do seguro, e resulta da adição de impostos e taxas legais ao prémio bruto.

Por outras palavras, o prémio de um seguro pode simplesmente ser dividido em duas grandes componentes: prémio puro e despesas adicionais. Assim sendo, surgem os prémios de inventário, comercial, bruto e total. Foquemo-nos então no prémio puro, o qual irá então corresponder, num sentido lato, ao valor atual esperado dos benefícios a prestar no futuro, possuindo portanto não só uma componente financeira (valor temporal do dinheiro), mas também uma componente probabilística (materialização de pagamentos futuros incertos). Neste sentido, o prémio puro é muitas vezes visto como sendo o produto entre a frequência com que estes ocorrem e o custo esperado a estes associado, o que faz sentido - estamos a multiplicar custos por sinistro por um número de sinistros por unidade de tempo, pelo que o resultado será um custo total expresso em unidades monetárias por unidade de tempo (euros por ano, por exemplo).

Se uma seguradora desejar atrair compradores, deverá cobrar, como visto há pouco, prémios considerados equitativos. No geral, é do interesse de todas as partes envolvidas (seguradoras, tomadores de seguros e sobretudo entidades de supervisão e regulação) o estabelecimento de prémios justos, isto é, que sejam do agrado de todas as partes envolvidas. Na atividade seguradora, o termo equidade significa que a dois indivíduos que paguem o mesmo prémio devem corresponder os mesmos valores atuais esperados de indemnizações. Em adição, de um modo geral, quanto maior o risco, maior o prémio.

Como as seguradoras recebem elevadas somas na forma de prémios, têm o dever de saber geri-los. Assim, a determinação de prémios puros deve ser feita com prudência, pois apesar de prémios baixos ajudarem a atrair clientes, estes também podem levar à ruína da seguradora. Desde que disponham de suficiente experiência ou conhecimento de eventos anteriores (leia-se, observações em conjuntos de dados históricos, por vezes extensos), as seguradoras podem, através de profissionais denominados de atuários, usar tais dados para obter montantes apropriados a cobrar aos tomadores de seguros.

Seguradoras necessitam então de ser capazes de avaliar riscos, de maneira a decidir se um dado indivíduo deve de facto ser coberto por um seguro, e a que preço, sendo que quanto mais precisa for esta informação, mais perto estará a seguradora de tomar decisões corretas; no sentido inverso, dados insuficientes ou defeituosos podem fazer com que tomadores de seguros sejam incorretamente classificados em termos dos riscos que representam.

Assim sendo, é natural que se procure a construção ou, pelo menos, a obtenção de conjuntos/bases de dados que contenham variáveis que ajudem a distinguir riscos e, portanto, a atingir estes objetivos, podendo o uso de tais variáveis constituir uma vantagem competitiva para a empresa/companhia de seguros (se forem de qualidade e não estiverem a ser utilizadas por nenhuma concorrente). Por isso, o cálculo de prémios com base nos riscos encoraja a inovação na atividade seguradora, tanto ao nível de preços (os quais se tornam mais apelativos) como ao nível de produtos (os quais passam a cobrir riscos anteriormente não

seguráveis), através, por exemplo, de tais fatores de tarifação sofisticados. Esta prática incentiva ainda a promoção de comportamentos responsáveis.

Por outras palavras, riscos são classificados e prémios definidos de acordo com uma série de variáveis preditivas, ou seja, fatores de tarifação. Tal classificação é baseada na premissa de que indivíduos no mesmo grupo vivenciam uma mortalidade ou sinistralidade largamente consistente. A análise dos dados históricos permite prever a probabilidade/frequência e a severidade associadas à materialização de riscos em cada um dos grupos de risco.

É também aqui que entra um princípio probabilístico-estatístico muito conhecido, a Lei dos Grandes Números, a qual é a pedra basilar da atividade seguradora. De uma forma simplista, esta lei afirma que à medida que o grupo de indivíduos ou bens seguros cresce, maior será o grau de confiança com que o impacto financeiro do risco para a seguradora pode ser calculado, isto é, mais precisas serão as previsões associadas a perdas. Por outras palavras, quantos mais contratos de seguro (de um mesmo tipo, de natureza similar e com indivíduos estatisticamente independentes entre si) forem celebrados por uma seguradora, regra geral, mais exatas serão as estimativas calculadas por esta. Importa referir que esta maior certeza da indemnização média por apólice não está acessível a meros indivíduos.

Com efeito, a seguradora assume que saberá com total certeza o valor monetário médio dos benefícios entregues no futuro, cobrando de seguida prémios que correspondam a este valor. Claramente, a seguradora sabe que não consegue prever de forma exata este valor médio, mas pode (como veremos nos princípios de cálculo de prémios, em anexo) considerar as perdas totais esperadas e o potencial de variação das mesmas para estimar tais prémios.

As seguradoras têm então em consideração muitos fatores no momento de estimação de prémios puros, os quais são determinados sobretudo através da análise de dados históricos em grupos homogéneos representando riscos semelhantes. Geralmente, quanto maior o número de fatores de risco utilizados para dividir os segurados em grupos mais pequenos, mais corretas serão as premissas associadas à frequência/severidade de um sinistro. No entanto, ao determinar o número de grupos de risco em que irão ser divididos os segurados é necessário encontrar um equilíbrio entre ter poucos grupos (em que os riscos não são homogéneos), e demasiados grupos (em que o número de segurados em cada grupo poderá ser demasiado pequeno para a análise ser estatisticamente significativa).

Temos, contudo, de prestar especial atenção ao facto de haver certos riscos sistemáticos ou não-diversificáveis, nos quais o pressuposto de independência que alimenta a Lei dos Grandes Números falha, pois afetam a maioria ou até todos os membros de um dado grupo em simultâneo. Por exemplo, um incêndio que destrói uma habitação fará com que seja mais provável a destruição (total ou parcial) de outras habitações e patrimónios na vizinhança desta, pelo que uma mesma seguradora não deve segurar todas as habitações, lojas e demais imóveis de uma dada área contra incêndios, nem contra outros fenómenos ou ocorrências, sob pena de colocar em questão a sua própria situação financeira.

É a Lei dos Grandes Números, aliada à existência de risco, que tornam o contrato de seguro viável e lhe conferem valor económico. Para que um seguro seja economicamente viável é portanto necessária, no geral, a existência de um vasto número de riscos independentes e, se possível, semelhantes, como visto na definição de risco segurável.

Devido à natureza competitiva do mercado de seguros, há ainda lugar a preocupações estratégicas no momento de definição de preços por parte das seguradoras. Por isso, o prémio final depende também da

estratégia de negócio de cada seguradora; por exemplo, de modo a ganhar quota de mercado, uma companhia pode desejar posicionar os seus produtos como os mais baratos no mercado, diminuindo as margens do lucro unitárias, compensando tal redução com um aumento nas vendas, efetivamente concorrendo na quantidade.

Principais ramos de seguros e modalidades mais comuns

De uma forma mais concreta, quando falamos de um contrato de seguro, podemos estar perante um:

- Seguro de vida, o qual visa realizar pagamentos a um ou mais beneficiários designados pelo detentor da apólice/tomador do seguro em caso da morte e/ou da sobrevivência da pessoa segura;
    - Podem também incluir coberturas complementares de incapacidade para o trabalho ou de invalidez;
    - Os mais conhecidos serão, porventura, os seguros de vida associados ao crédito habitação (SVCH), os quais visam liquidar a porção do crédito ainda em dívida ao banco, no caso de morte da pessoa segura;
- Seguro *unit-linked*, o qual é um seguro ligado a fundos de investimento, sendo por isso mais arriscado para o tomador do seguro;
- Planos poupança-reforma (PPR), os quais incentivam à poupança de longo prazo e ao complemento das pensões concedidas pelo Estado através da Segurança Social;
- Seguro de acidentes pessoais, o qual visa substituir ou cobrir, no todo ou em parte, o rendimento do beneficiário em caso de lesão corporal, invalidez temporária ou permanente, ou morte da pessoa segura, por causa súbita, externa e imprevisível;
- Seguro de acidentes de trabalho, o qual tem caráter obrigatório para todos os trabalhadores e visa ajudar no pagamento de despesas médicas resultantes de um acidente de trabalho, bem como de eventuais pensões por morte ou incapacidade permanente que se justifiquem;
- Seguro automóvel, o qual visa cobrir o risco de ocorrência de um acidente automóvel;
- Seguro de incêndios, os quais visam cobrir danos causados por incêndio e são obrigatórios para imóveis em propriedade horizontal;
- **Seguro multirriscos habitação**, os quais englobam seguros contra incêndios, uma vez que também cobrem danos causados por incêndio e, possivelmente, danos causados por fenómenos como inundações, tempestades, danos por água, fenómenos sísmicos, riscos elétricos, deslizamentos de terra, tornados e roubos (entre outros);
- Seguro de responsabilidade civil (RC), o qual tem o propósito de cobrir o risco de o segurado ter de indemnizar outrem por danos que lhe cause;
- Seguro de saúde, no qual a companhia de seguros cobre riscos relacionados com a saúde dos beneficiários, através do pagamento total ou parcial de despesas relacionadas com a prestação de cuidados médicos (como o recurso a dentistas).

De uma forma mais abrangente, estes seguros podem pertencer:

- Ao ramo Vida, se dependerem diretamente da vida ou morte da pessoa segura;
- Aos ramos Não Vida, caso contrário.

Dentro dos ramos não vida, podemos ainda distinguir os ramos associados a riscos pessoais dos ramos associados a riscos patrimoniais.

Do ramo vida fazem parte essencialmente (e sem surpresas) os seguros de vida, bem como os seguros de nupcialidade/natalidade, os seguros ligados a fundos de investimento (*unit linked*) e as operações de capitalização (de acordo com a ASF). Por isso, os três primeiros tipos de seguros em cima mencionados pertencem ao ramo Vida, estando os restantes tipos inseridos em ramos Não Vida, quer tenham natureza pessoal (como o seguro de acidentes pessoais, o seguro de acidentes de trabalho e o seguro de saúde) ou patrimonial (como o seguro de incêndios e o seguro multirriscos habitação).

Legislação relevante. Supervisão, controlo prudencial e regulação de companhias de seguros

A atividade seguradora, enquanto atividade económica, é afetada pela legislação em vigor em cada local e momento. Neste sentido, será indicada de seguida a legislação considerada, à data de escrita deste documento, mais relevante para a atividade seguradora em Portugal.

Podemos começar pelo regime de Solvência II, muito relevante para a atividade seguradora à escala europeia, o qual surge através da Diretiva n.º 2009/138/CE, publicada em 25 de novembro de 2009, a qual foi objeto de alterações posteriores. Este regime visa, essencialmente:

- Fixar requisitos de capitais a cumprir por cada seguradora ou resseguradora, de modo a afastar ao máximo possível um cenário de insolvência, sendo tais capitais definidos em função dos riscos que estas enfrentam;
- Padronizar a regulação exercida ao nível de seguros em toda a União Europeia, efetivamente permitindo a criação de um mercado único.

A diretiva responsável pela entrada em vigor do regime de Solvência II foi transposta para o ordenamento jurídico português através da Lei n.º 147/2015, publicada a 9 de setembro de 2015 no Diário da República n.º 176/2015, Série I, a qual é usualmente denominada de Regime Jurídico de Acesso e Exercício da Atividade Seguradora e Resseguradora ou, abreviadamente, RJASR. Esta lei visa por sua vez alterar o Decreto-Lei n.º 72/2008, divulgado no Diário da República n.º 75/2008, Série I de 16 de abril de 2008, a qual estabelece o regime jurídico do contrato de seguro. Estes dois regimes são considerados diplomas base, isto é, são essenciais para o enquadramento jurídico da atividade seguradora.

Devido à sua relevância e natureza, certos seguros em vigor na ordem jurídica portuguesa podem ter, em alguns contextos e atividades profissionais específicas, caráter obrigatório. É isto o que sucede nos seguros de incêndios, em particular em edifícios em propriedade horizontal, estando tal necessidade prevista:

- No Código Civil (na redação do Decreto-Lei n.º 267/94, de 25 de Outubro, o qual altera o Artigo 1429º do CC);
- No Artigo 5º do Decreto-Lei n.º 268/94, de 25 de outubro;
- Na Norma n.º 18/2000-R, de 21 de dezembro, alterada pela Norma n.º 13/2005-R, de 18 de novembro.

A título adicional, a atividade seguradora tem também de cumprir com outras leis de caráter mais geral – como leis que visem combater a discriminação baseada em sexo/género, ou leis de defesa dos direitos do consumidor.

O setor segurador está também sujeito à supervisão e regulação de entidades dedicadas para o efeito. A Autoridade de Supervisão de Seguros e Fundos de Pensões (ASF) é a autoridade nacional responsável pela regulação e supervisão de companhias de seguros, fundos de pensões, mediadores de seguros e demais entidades com atividades relacionadas, garantindo o cumprimento das leis e normas aplicáveis, com o intuito de incentivar condutas corretas nas mesmas.

A missão da ASF é a de garantir que estas entidades se comportam de forma diligente, equitativa e transparente no seu relacionamento com tomadores de seguros, segurados, beneficiários e lesados (cujos direitos protege), assim assegurando o bom funcionamento do mercado segurador e dos fundos de pensões.

É a ASF que concede permissão para a constituição e existência de companhias de seguros (ou de resseguros), através da emissão de uma autorização para operar em solo nacional.

Existe ainda, ao nível comunitário, uma outra autoridade responsável pela supervisão do setor segurador e dos fundos de pensões - a Autoridade Europeia dos Seguros e Pensões Complementares de Reforma (EIOPA - *European Insurance and Occupational Pensions Authority*), a qual surge a 1 de janeiro de 2011, em substituição do Comité Europeu de Supervisão de Seguros e Pensões Complementares de Reforma (CEIOPS – *Committee of European Insurance and Occupational Pensions Supervisors*), através do Regulamento (UE) 1094/2010. A principal missão da EIOPA é a de proteger o interesse público (sobretudo o dos consumidores), possuindo esta a responsabilidade de contribuir para a transparência, confiança e estabilidade no setor financeiro no curto, médio e longo prazo.

A função atuarial e a profissão de atuário

Seguradoras desejam aceitar riscos provenientes de outros agentes económicos de uma forma prudente e adequada à sua realidade, para garantir tanto o pagamento de benefícios a detentores de apólices como a obtenção de rentabilidades razoáveis para os *shareholders*, e ainda cumprir uma função social de importância, ao mitigar consequências financeiras de eventos adversos. Enquanto entidades inseridas no sistema financeiro global, estas estão expostas a riscos decorrentes da conjuntura económica e da natural volatilidade dos mercados financeiros.

Por tudo isto, o setor segurador exige especial atenção, sobretudo devido à prudência necessária na estimação adequada de responsabilidades (a qual exerce influências na gestão de ativos e de passivos), tendo em vista a sustentabilidade da seguradora no médio e no longo prazo. Por outro lado, cada vez mais a realidade do mundo que nos rodeia e que engloba as seguradoras se torna mais complexo e, por isso, mais difícil de modelar.

Pela sua formação e pelas aptidões que possuem na identificação, mensuração e gestão de riscos, atuários têm desde cedo assumido um papel único ao nível da sustentabilidade financeira de companhias de seguros.

De uma forma ampla, os atuários são os profissionais capazes de modelar fenómenos aleatórios através do recurso a modelos quantitativos, e assim ajudar as organizações que a eles recorrem a perceber a

quais riscos estão expostas, bem como quais os potenciais custos daí decorrentes. Com base nas recomendações de atuários, estas organizações procuram analisar as consequências financeiras destes riscos e obter garantias de solvência, isto é, procuram viabilizar a realização de pagamentos a quem tem direito aos mesmos.

Por isso, um atuário deve ser capaz de, num contexto de elevada complexidade:

- Analisar dados;
- Avaliar e gerir riscos financeiros;
- Comunicar tais informações a pessoas que não sejam especialistas fornecendo, portanto, conselhos sobretudo de natureza financeira (mas também de natureza comercial e estratégica);

A intervenção dos atuários – dos quais se espera uma atuação independente, rigorosa, idónea e transparente – é, portanto, indispensável para assegurar estabilidade e confiança no setor segurador, sobretudo tendo em conta a recente implementação e desenvolvimento de regimes como o Solvência II, os quais geram não só desafios, mas também oportunidades.

Muitas vezes, atuários são tradicionalmente empregues por seguradoras (tanto de ramos vida como não vida), resseguradoras e/ou fundos de pensões. No entanto, atuários também são procurados por entidades como empresas de consultadoria, entidades de supervisão e regulação, entidades de gestão de investimentos ou de riscos financeiros (como bancos e agências de *rating*), entidades governamentais e, por fim, empresas (sobretudo em departamentos responsáveis por assuntos remuneratórios), podendo ainda exercer a sua profissão em regime *freelance*/por conta própria.

Numa seguradora, a principal tarefa de um atuário é a de assegurar que as somas obtidas através da cobrança de prémios e dos rendimentos de investimentos são suficientes para cobrir o pagamento de benefícios – se isto não acontecer, a seguradora irá incorrer em situações de incumprimento obviamente injustas. O principal desafio do atuário é lidar com a aleatoriedade que envolve este problema e que tem diversas fontes – pois desconhecemos *à priori*:

1. Quantos benefícios terão de ser pagos (em número);
2. Quais os montantes (valores monetários) que terão de ser pagos em cada benefício;
3. Quais serão os rendimentos oriundos dos investimentos efetuados.

Como esperado, o atuário faz uso de métodos probabilísticos e estatísticos para lidar com tal aleatoriedade. A missão do atuário é então a de resolver problemas de tarifação[2] de seguros (*pricing*), de constituição de provisões[3] técnicas (*reserving*) ou de modelação de capitais[4] para fazer face a riscos mais gerais (*capital modelling*), devendo para estes efeitos:

- Construir modelos para uso futuro (e estimar os seus parâmetros);
- Extrair conclusões através destes modelos;
- Indicar possíveis desfechos futuros, sobretudo ao nível de solvência e de rentabilidade.

[2] Alocando despesas a apólices, determinando margens de segurança e de lucro apropriadas, e definindo ainda elementos como termos de resgate, de redução e de transferência de apólices (no ramo vida).
[3] Estas correspondem somente às responsabilidades futuras esperadas para a seguradora, ou seja, apenas a passivos, resultando da projeção de perdas/pagamentos de benefícios futuros.
[4] Os quais se distinguem de reservas pois podem compreender, por exemplo, correções nos mercados financeiros, nos quais a seguradora pode ter investimentos – e correspondendo, portanto, tanto a ativos como a passivos.

A função atuarial é geralmente reconhecida como sendo uma das quatro principais funções de controlo dentro de cada seguradora, pertencendo à segunda linha de defesa do modelo/sistema de governação, sendo essencial para um vasto conjunto de *stakeholders*, como órgãos de direção, supervisores, auditores, entre outros. A função atuarial deve ainda ser independente face a elementos de gestão e a atividades operacionais, sendo que esta independência permite a *stakeholders* como entidades de supervisão confiar que os sistemas de controlo da seguradora estão em bom estado, sendo esta seguradora alvo de uma supervisão externa menos intrusiva.

De acordo com o Artigo 48.º da Diretiva associada ao regime de Solvência II, as responsabilidades previstas para a função atuarial de uma seguradora passam tipicamente, e a um nível primário, pela comunicação de opiniões, recomendações e resultados a órgãos de administração da companhia, de forma periódica.

Adicionalmente, um atuário deve manter a sua formação atualizada, uma vez que a profissão que exerce é alvo da influência de inovações tecnológicas – as quais trazem não só oportunidades, mas também desafios, como a existência de ataques informáticos, os quais motivam o surgimento de seguros para riscos cibernéticos (*cyberrisks*) que, para serem comercializados, precisam de ser precificados por atuários - o que requer alguma criatividade, sobretudo em termos de modelação, dado serem relativamente recentes.

Sobre a Aegon Santander Portugal (ASP)

Sobre a AEGON e o Banco Santander

O Banco Santander foi criado em 1857, na província de Santander, em Espanha; posteriormente, estendeu-se para todo o país através da compra de vários bancos, entrando na década de 1980 em Portugal. Em 2000 dá-se a compra do grupo financeiro Totta e Açores, passando a denominar-se Banco Santander Totta, e em 2015 é adquirido o BANIF (na sequência da crise no mesmo). O Banco Santander Totta é também mediador da Santander Totta Seguros, seguradora que comercializa os seus seguros nos balcões do Banco.

Por sua vez, a Aegon surge em 1844, na Holanda, tendo crescido através da fusão e aquisição de diversas pequenas seguradoras especializadas nas múltiplas áreas desta atividade. Em 1983 dá-se a fusão das holandesas Ago e Ennia e muda o nome para AEGON. É também a partir de 1983 que se dá a grande expansão da empresa na América, Europa e Ásia.

Presente no mercado português desde 2014, a Aegon Santander Portugal nasceu precisamente da aliança entre a Aegon e a Santander Totta Seguros, oferecendo assim uma gama alargada, especializada e de qualidade de Seguros, através de duas companhias (Aegon Santander Portugal Vida e Aegon Santander Portugal Não Vida), comercializados em exclusivo através do Grupo Santander em todo o território português.

Sobre a Aegon Santander Portugal

De forma mais detalhada, a Aegon Santander Portugal (doravante, apenas ASP) é uma entidade fundada no final de 2014 e composta por duas companhias de seguros (as quais partilham estruturas):

- A Aegon Santander Portugal Vida (doravante, ASP Vida);
- A Aegon Santander Portugal Não Vida (doravante, ASP Não Vida);

sendo o resultado de uma parceria entre o Grupo Aegon e o Grupo Santander, os quais determinam a visão estratégica da ASP.

Em particular, tanto na ASP Vida como na ASP Não Vida, verifica-se que 51% do capital é detido pela Aegon Spain Holding B.V., sendo os restantes 49% detidos pela Santander Totta Seguros, S.A., sendo a comercialização de seguros feita através do Banco Santander Totta.

A missão da ASP é a de proteger a vida, a família, a saúde e os bens das famílias e empresas que nela depositam confiança, representando, portanto, estes clientes o seu principal foco. Esta proteção é oferecida através de produtos como o **Proteção Lar**, seguro multirriscos habitação que serve para proteger um imóvel (edifício) e/ou os bens nele existentes (recheio) contra imprevistos adversos como furto, inundações, incêndio.

Adicionalmente, o bem-estar dos seus colaboradores é também uma prioridade, tendo a ASP sido considerada (desde 2016) pela revista Exame como sendo uma das 100 melhores empresas para trabalhar em Portugal.

Ao longo do tempo a ASP tem crescido em indicadores relevantes como o número de apólices e capitais seguros, bem como em prémios recebidos e benefícios prestados.

A ASP visa cumprir todos os seus compromissos e recompensar todos os seus *stakeholders*, privilegiando tanto o rigor e a transparência como a criação de valor para os seus *shareholders*, sem descurar a sua sustentabilidade financeira nem a capacidade de aproveitar oportunidades de investir e de crescer.

Contexto económico e social

Importa referir a realidade macroeconómica no qual este trabalho se insere e foi elaborado, dado o mesmo não existir fora deste contexto.

Para potenciar um crescimento económico equilibrado e um elevado nível de emprego, nos últimos anos (mais concretamente, desde inícios de 2015), o Banco Central Europeu (BCE) tem procurado estimular a economia através de programas de *quantitative easing*, os quais envolvem a adquisição de ativos em grande escala por parte do mesmo.

Ao adquirir grandes volumes de títulos de dívida emitidos por instituições da Zona Euro, o BCE aumentou a oferta[5] de financiamento, aumento esse que provoca uma deslocação para o exterior na curva de oferta, algo que resulta, *ceteris paribus*, em descidas no preço do dinheiro – ou seja, nas taxas de juro. Dada a natureza muitas vezes soberana e por isso extremamente segura desta dívida, estas taxas de juro servem de referência para outras, pelo que esta redução motiva diminuições em muitas outras taxas de juro praticadas nas economias europeias.

Menores taxas de juro tornam o acesso a financiamento mais acessível tanto para indivíduos, os quais podem mais facilmente financiar o consumo de bens e serviços, como para empresas, cujos menores custos de capital fazem com que alguns projetos de investimento passem a ser considerados viáveis, sendo que a execução dos mesmos leva à contratação de trabalhadores – os quais passam a ter mais rendimentos e menor desemprego – e à produção de mais bens e serviços. Tudo isto fomenta o crescimento económico.

Apesar de políticas de *quantitative easing* terem sido inicialmente consideradas medidas de caráter temporário, a verdade é que as mesmas nunca mais foram verdadeiramente "desativadas" pelo BCE. Adicionalmente, a pandemia do covid-19 veio fazer com que a Reserva Federal estado-unidense – o *Fed* – adotasse também políticas de *quantitative easing*.

Como seguradoras investem maioritariamente em títulos de dívida como obrigações, taxas de juro baixas constituem uma dificuldade a ultrapassar. Por isso, podemos afirmar que o setor segurador é dos que mais sofre com a existência destas taxas, no sentido em que estas trazem às seguradoras maiores custos no presente e menores rendimentos no futuro.

Companhias de seguros podem investir os prémios que arrecadam em obrigações detidas até à maturidade. Quando isto acontece, cortes nas taxas de juro fazem com que estes títulos sejam substituídos

[5] Dado que quem realiza investimentos recebe juros, e é assim considerado "vendedor".

na sua maturidade por outros similares (para manter as proporções entre títulos de dívida e restantes ativos), mas com menores remunerações, o que terá impactos negativos nos resultados da seguradora. À medida que o tempo passa, um número crescente de obrigações atingem a maturidade e forçam as seguradoras (as quais têm uma forte necessidade de investir de novo estes capitais) a aplicar fundos em obrigações mais recentes e menos rentáveis.

Assumindo ainda que os montantes reportados, nos balanços financeiros, das responsabilidades futuras esperadas correspondem a valores atuais, uma descida na taxa a que estes fluxos são atualizados motiva um aumento no valor destes passivos, o qual é registado em demonstrações de resultados como um custo que, *ceteris paribus*, faz com que os resultados da companhia diminuam.

Porém, ao contrário do que acontece sobretudo no ramo Vida, seguradoras de ramos Não Vida (como o ramo de seguros patrimoniais, ao qual pertence o seguro de multirriscos habitação em estudo neste trabalho) tendem a não ser tão afetadas por taxas de juro baixas, uma vez que possuem passivos de curto e médio-prazo (na maioria, inferior a 3 anos).

Adicionalmente ao estimularem o crescimento económico, medidas de *quantitative easing* podem até trazer benefícios interessantes para os ramos Não Vida. De acordo com José Galamba de Oliveira, presidente da Associação Portuguesa de Seguradores, estes ramos são tradicionalmente mais dependentes da evolução da atividade económica, sendo que com o crescimento económico registado em Portugal nos últimos anos, deu-se também uma expansão destes ramos. Tal não surpreende; nos seguros multirriscos habitação em particular, este crescimento pode motivar um aumento no número de imóveis vendidos e, por extensão, no número de contratos de seguro deste tipo.

Torna-se então necessário estudar os efeitos da pandemia covid-19, a qual arrastou muitas economias para uma situação de recessão. Ora, o seguro tende a ser um produto no qual diminuições ao nível de rendimentos podem causar impactos desproporcionalmente negativos nas quantidades de seguros adquiridos, pelo que se espera que esta pandemia tenha impactos muito pouco desejáveis para o setor segurador.

Ao nível de seguros em vigor durante a pandemia, é de esperar um impacto mais direto em seguros de saúde (dado o vírus causar doenças) e até de vida (dado estas doença poderem, no limite, resultar na morte). No seguro de multirriscos habitação, há um menor impacto direto, mas um maior impacto indireto, pois numa recessão a compra de habitações e consequente contratação deste tipo de seguros tende a abrandar ou até estagnar. Já ao nível de seguros de desemprego, o covid-19 irá provavelmente causar uma crise responsável pelo pagamento de uma quantidade anómala de benefícios. Estes pagamentos, sendo aleatórios, não serão propriamente independentes – antes fortemente correlacionados, dado esta crise poder afetar uma parte significativa da população.

Sobre os seguros multirriscos habitação

Seguros de multirriscos habitação

Seguros impulsionam a atividade económica e são impulsionados por esta, a qual por sua vez está invariavelmente relacionada com o consumo feito pelos agentes económicos. Regra geral, uma das transações mais importantes na vida de uma pessoa prende-se com a compra de uma habitação à qual possa chamar de lar.

A adquisição de uma habitação exige, como garantia contra riscos, a assinatura de um seguro multirriscos habitação (abreviadamente, seguro MrH), o qual pode englobar, segundo a ASF, proteções ao nível:

- Da reparação de danos causados na própria fração ou noutras frações do edifício, por ocorrência de riscos distintos do incêndio como, por exemplo, inundações, tempestades, raios, explosões, riscos elétricos, aluimentos de terras, e demais catástrofes naturais/fenómenos meteorológicos extremos;
- Da reparação de danos causados nos bens móveis da habitação (recheio);
- Da indemnização por furto qualificado ou roubo de bens de uso pessoal;
- Da responsabilidade civil do segurado e pessoas do seu agregado familiar (caso seja necessário indemnizar terceiros por danos involuntariamente causados, dos quais resultem lesões materiais e ou corporais).
- De indemnizações por morte do segurado ou cônjuge, em consequência de incêndio, queda de raio, explosão ou roubo, quando ocorrida na habitação;

bem como outras coberturas, de caráter complementar.

Em bom rigor, o seguro MrH não é obrigatório, pois o seguro de incêndio é o único obrigatório por lei para edifícios em regime de propriedade horizontal (como, por exemplo, complexos habitacionais constituídos por apartamentos), cobrindo o risco de danos provocados no imóvel por incêndio, e devendo, de acordo com a ASF, cobrir cada fração autónoma e as partes comuns do edifício (telhado, escadas, elevadores, garagem, etc.).

Porém, como coberturas para fazer face a incêndios estão também incluídas em seguros multirriscos habitação, estes últimos são mais abrangentes, sendo habitualmente sugeridos ou, até, exigidos[6] por bancos, empresas de gestão de condomínio e companhias de seguros; assim, o proprietário consegue proteger o seu imóvel contra uma panóplia de azares que nele possam ocorrer. De facto, o seguro MrH tende a ser o mais escolhido em Portugal, uma vez que proporciona mais garantias do que o seguro de incêndio, a um preço não muito superior (na maioria dos casos).

[6] Atitude que se compreende: salvaguardar o imóvel é do interesse destas entidades, enquanto credoras.

Como os montantes que a seguradora terá de pagar, em caso de sinistro, irão depender da natureza do mesmo e do seu enquadramento, constata-se que o prémio associado a um seguro MrH varia de acordo com o conjunto de coberturas subscritas.

Seguros MrH compreendem usualmente, como descrito há pouco, conjuntos de coberturas pré-determinadas, sendo possível agregar coberturas de caráter complementar. Os produtos disponibilizados pelas companhias de seguros tendem a ser particionados em três níveis - básico, intermédio e completo - com maiores prémios, mas também coberturas mais numerosas e abrangentes. Sendo Portugal um país com elevado risco sísmico[7] - sobretudo em Lisboa, no Algarve e nos Açores - é essencial considerar a aquisição de coberturas para fenómenos sísmicos.

Todos os contratos de seguros multirriscos têm como referência um dado montante a pagar, denominado de capital seguro, o qual corresponde ao maior pagamento que a seguradora pode ser obrigada a realizar devido a um sinistro enquadrado.

Quem adquire uma habitação para nela viver nem sempre se preocupa apenas com o estado do imóvel em si. Por isso, um seguro multirriscos habitação visa, após contratado, proteger o proprietário de um imóvel de danos ocorridos na sua estrutura – o edifício – ou nos bens existentes no interior desta estrutura – conteúdos usualmente denominados de recheio – sendo ainda possível contratar ambas as proteções em simultâneo.

Portanto, no momento de subscrição de um seguro MrH, é importante uma correta avaliação do valor dos bens a proteger (sendo esta responsabilidade do segurado), de maneira a que o capital seguro cubra na totalidade:

- O custo de reconstrução do imóvel (caso a cobertura para edifícios seja contratada);
- O custo de substituição dos conteúdos (caso a cobertura para recheios seja contratada);

Seguros multirriscos habitação estão sujeitos à regra da proporcionalidade, explicada de seguida.

Exemplo: *Se o proprietário de um imóvel cujo valor de reconstrução é de 100000 euros decidir assinar um contrato de seguro no qual estipula um capital seguro de 70000 euros, considera-se que apenas 70% do valor do edifício está coberto pelo seguro. Se ocorrer neste imóvel um sinistro (enquadrado) avaliado em 40000 euros, apesar do capital seguro exceder esse valor, a seguradora só pagará 70% das perdas - ou seja, 40000 × (70%) = 28000 euros, ficando os restantes 40000 − 28000 = 12000 euros a cargo do tomador.*

Este princípio encoraja os tomadores de seguro a procurar uma fiel avaliação do património que estão a segurar. Também não faz sentido definir um capital seguro superior ao valor de mercado do imóvel, uma vez que a companhia de seguros só indemnizará o real valor dos bens danificados, de maneira a evitar que a materialização de um risco seja lucrativa para o beneficiário.

Quaisquer obras de melhoria ou outras alterações ao edifício devem também ser acompanhadas do ajuste correspondente ao capital seguro, de maneira a continuar a cobrir o valor do imóvel na sua totalidade, sob pena do seguro agora desatualizado vir a cobrir apenas o valor económico anteriormente existente, e não o valor gerado após tais melhorias.

[7] Quer pela probabilidade de ocorrência de sismos, quer pelo valor da destruição que estes provocariam no património.

Gestão de sinistros em seguros multirriscos habitação

Estando o pagamento de sinistros e até a própria definição de prémios a pagar dependente da ocorrência e da severidade de sinistros, chega-se à conclusão de que o sinistro é também um conceito cujo estudo é fundamental, mesmo no contexto desta tese.

Assim, torna-se necessário saber de que forma é que uma companhia de seguros lida com pedidos de indemnização associados a sinistros, sendo que dependendo do ramo, e até da modalidade, o processo de gestão de sinistros pode sofrer algumas alterações.

No geral, e como visto na Unidade Curricular de Atividade Seguradora, o processo de gestão de sinistros em seguros MrH (e noutros seguros patrimoniais) é o seguinte:

1. Ocorrência do sinistro;
2. Participação do sinistro;
3. Abertura do processo associado ao sinistro;
4. Avaliação dos danos causados pelo sinistro;
5. Possível investigação dos danos causados pelo sinistro – por exemplo, peritagens;
6. Reparação dos danos;
7. Pagamento de indemnizações e de serviços;
8. Encerramento do processo;
9. Possível reabertura do processo (o qual recomeça numa das duas fases anteriores a esta).

As fases do processo de gestão de sinistro e o tempo decorrido entre a ocorrência e o encerramento dependem de diversos fatores. Em particular, no ramo Patrimoniais, no qual se insere o seguro MrH, sinistros tendem a ser processados de forma ágil, dando origem a pagamentos mais céleres (ao contrário, por exemplo, do ramo Pessoais, no qual não raras vezes é exigido o recurso a tribunais, podendo perfeitamente haver lugar ao pagamento de pensões até ao fim da vida do sinistrado).

Teoria dos Modelos Lineares Generalizados

A família exponencial de distribuições

Para percebermos o que se entende por modelo linear generalizado, temos de apresentar primeiro a noção de família exponencial de distribuições. Iremos então definir esta família através das propriedades partilhadas pelas distribuições que a ela pertencem.

Sejam Y_i ($i \in \{1,2,\dots,n\}$) variáveis aleatórias independentes com distribuições comuns (mas parâmetros possivelmente distintos). Sizemos que a distribuição de Y_i pertence à família exponencial de distribuições se a função massa de probabilidade ou a função de densidade de probabilidade de Y_i (consoante esta seja uma variável aleatória discreta ou absolutamente contínua, respetivamente) puder ser escrita na forma

$$f_{Y_i}(y;\theta_i,\varphi_i) = exp\left(\frac{y\theta_i - b(\theta_i)}{a(\varphi_i)} + c(y,\varphi_i)\right)$$

onde:

- θ_i é um parâmetro canónico de localização que é função de $\mu_i = E(Y_i)$;
- φ_i é um parâmetro de dispersão positivo ($\varphi_i > 0$), o qual é igual a uma constante em certas distribuições;
- $a(\varphi_i)$ é uma função real de variável real positiva e contínua, a qual assume habitualmente a forma φ/w_i;
- $b(\theta_i)$ é uma função real de variável real de classe C^2;
- $c(y,\varphi_i)$ é uma função real a duas variáveis reais;

A vantagem de expressar diversas distribuições no formato típico da família exponencial prende-se com o facto das propriedades válidas para esta família serem depois aplicáveis numa variedade de casos particulares. Por exemplo, podemos demonstrar que:

$$\mu_i = b'(\theta_i)$$

ou seja, θ_i é função de μ_i, e que:

$$V(\mu_i) = b''(\theta_i)a(\varphi_i)$$

ou seja, a função variância $V(\mu_i)$ indica-nos que alterações no valor médio μ_i podem conduzir a alterações na variância de Y_i.

É possível, após manipulações algébricas, demonstrar que as distribuições Normal, Binomial, de Poisson, Gama e Inversa Gaussiana pertencem à Família Exponencial.

Formulação do modelo linear generalizado (MLG) e casos particulares

O modelo clássico de regressão linear. Transformações e alertas a ter em mente

O conceito de modelo linear assume uma natureza central em Probabilidade e em Estatística. O modelo clássico de regressão linear é usualmente caraterizado pelas suas propriedades essenciais:

- Independência: as variáveis aleatórias $Y_1, Y_2, \ldots, Y_n$ são consideradas independentes;
- Linearidade: considera-se que o valor esperado $\mathbb{E}\left(\mathrm{Y_i}\middle|\vec{\mathrm{X}}_\mathrm{i} = \vec{\mathrm{x}}_\mathrm{i}\right)$ pode, para cada indivíduo/unidade estatística *i*, ser expresso enquanto combinação linear de parâmetros associados às variáveis explicativas/preditivas $\vec{\mathrm{X}}_\mathrm{i} = \left(X_{i1}, X_{i2}, \ldots, X_{ip}\right)$, ou seja,

$$\hat{Y}_i = \mathbb{E}\left(\mathrm{Y_i}\middle|\vec{\mathrm{X}}_\mathrm{i} = \vec{\mathrm{x}}_\mathrm{i}\right) = \beta_1 \mathrm{x_{i1}} + \ldots + \beta_\mathrm{p} \mathrm{x_{ip}}$$

$$Y_i = \beta_1 \mathrm{x_{i1}} + \ldots + \beta_\mathrm{p} \mathrm{x_{ip}} + \varepsilon_i$$

- Normalidade[8]: considera-se que $(\mathrm{Y_i}|\vec{\mathrm{x}}_\mathrm{i}) \cap \mathcal{N}(\mu_i, \sigma^2)$, ou seja, $\varepsilon_i \cap \mathcal{N}(0, \sigma^2)$;
- Homoscedasticidade: considera-se que $var(\mathrm{Y_i}|\vec{\mathrm{x}}_\mathrm{i}) = var(\varepsilon_i) = \sigma^2$, isto é, a variância de $(\mathrm{Y_i}|\vec{\mathrm{x}}_\mathrm{i})$ é constante.

Modelos clássicos de regressão linear são ajustados de acordo com o método dos mínimos quadrados, através do qual procuramos encontrar o vetor de estimadores $\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_\mathrm{p}\right)$ que minimiza o somatório de quadrados dos erros, isto é, que é solução do seguinte problema de otimização:

$$\min_{\vec{\beta} \in \mathbb{R}^n} SSE\left(\vec{\beta}\right) = \min_{\vec{\beta} \in \mathbb{R}^n} \sum_{i=1}^{n} e_i^2 = \min_{\vec{\beta} \in \mathbb{R}^n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{\vec{\beta} \in \mathbb{R}^n} \sum_{i=1}^{n} \left(y_i - \left(\beta_1 \mathrm{x_{i1}} + \ldots + \beta_\mathrm{p} \mathrm{x_{ip}}\right)\right)^2$$

Repare-se que tiramos partido de termos observado, em $\vec{\mathrm{X}}_\mathrm{i}$ e Y_i, os valores $\vec{\mathrm{x}}_\mathrm{i}$ e y_i para estimar estes coeficientes, isto é, para *aprender* o modelo.

Na verdade, e tendo ao nosso dispor n observações e p preditores, o que o modelo clássico de regressão linear múltipla estipula é o seguinte:

$$\begin{cases} y_1 = \beta_1 \mathrm{x_{11}} + \beta_2 \mathrm{x_{12}} + \ldots + \beta_\mathrm{p} \mathrm{x_{1p}} \\ y_2 = \beta_1 \mathrm{x_{21}} + \beta_2 \mathrm{x_{22}} + \ldots + \beta_\mathrm{p} \mathrm{x_{2p}} \\ \ldots \\ y_n = \beta_1 \mathrm{x_{n1}} + \beta_2 \mathrm{x_{n2}} + \ldots + \beta_\mathrm{p} \mathrm{x_{np}} \end{cases}$$

Utilizando notação matricial, ter-se-á $\vec{\boldsymbol{y}} = \boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}$, com:

$$\vec{\boldsymbol{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix}$$

[8] Apesar de ser útil para a realização de inferências (testes de hipóteses e intervalos de confiança/previsão), esta condição não faz parte da definição do modelo linear. Porém, se a dimensão da amostra for elevada e se as covariáveis apresentarem uma elevada variabilidade, os testes de hipóteses e intervalos de confiança desenvolvidos sob o pressuposto de normalidade são aplicáveis, com resultados aproximados, ao caso do modelo linear com resíduos não-normais. Ver, por exemplo, Sen e Srivastava (1990).

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\vec{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_p \end{bmatrix}$$

$$\vec{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

Podemos (assumindo que a matriz $\boldsymbol{X^T X}$ admite inversa) demonstrar que, em notação matricial, os estimadores de mínimos quadrados são dados por:

$$\vec{\boldsymbol{\beta}}_{OLS} = \left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{X^T} \vec{\boldsymbol{y}}$$

No caso de os erros seguirem uma distribuição Normal, é possível demonstrar que o método dos mínimos quadrados e o método da máxima verosimilhança conduzem à mesma solução.

Definam-se, respetivamente, a soma dos quadrados totais (*total sum of squares* - SST), a soma dos quadrados associados à regressão (*regression sum of squares* - $SSReg$) e a soma dos quadrados dos erros (*error sum of squares* - SSE) da seguinte forma:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSReg = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Podemos demonstrar que, num modelo de regressão linear múltipla no qual exista um termo constante, a seguinte decomposição é válida:

$$SST = SSReg + SSE$$

A existência de termo constante, nos modelos atuariais que iremos considerar na secção seguinte, fará com que a primeira covariável associada a cada observação seja sempre unitária, isto é, $x_{i1} = 1, \forall i \in \{1,2,\dots,n\}$. Ou seja, nestas condições, a primeira coluna de $\boldsymbol{X}$ será unitária.

O nosso interesse é o de construir modelos parcimoniosos, isto é, obter "os modelos mais simples que melhor se ajustem aos dados". Por isso, sendo p o número de parâmetros do modelo, para avaliar a qualidade de ajustamento de um modelo de regressão linear, devemos recorrer:

- Ao coeficiente de determinação, $R^2 \in [0,1]$:

$$R^2 = \frac{SSReg}{SST} = 1 - \frac{SSE}{SST}$$

- Este coeficiente mede a proporção de variabilidade dos dados que pode ser explicada pelo modelo de regressão linear;
- Mais uma vez, vemos que quanto maior a SSE, menor a qualidade de ajustamento do modelo;

- Ao coeficiente de determinação ajustado, $R^2_{ADJ} \in [0,1]$:

$$R^2_{ADJ} = 1 - \frac{n-1}{n-p}(1-R^2) \leq R^2$$

- Esta é uma variante do coeficiente de determinação que penaliza a introdução de variáveis no modelo de regressão linear;
- Este indicador castiga/penaliza modelos mais complexos;

- Ao critério de informação de Akaike (AIC):

$$AIC = \underbrace{-2ln\left(\mathcal{L}(\vec{\boldsymbol{\beta}}, \sigma^2)\right)}_{afere\ o\ ajustamento\ do\ modelo} + \underbrace{2p}_{reflete\ o\ n^{o}.de\ preditores}$$

- Quanto maior a log-verosimilhança do modelo, *ceteris paribus*, menor (mais negativo) será o valor do AIC;
- Quanto menor for o número de variáveis explicativas e/ou preditivas, *ceteris paribus*, mais negativo será o valor do AIC;
- Como tal, quanto menor for o valor do AIC, maior a parcimónia do modelo (e vice-versa).

Podemos testar estatisticamente se os coeficientes β_j ($j \in \{1,2,\ldots,p\}$) são significativamente diferentes (ou não) de um dado valor $\beta_{j|H_0}$ (usualmente 0) – por exemplo, testar:

$$H_0: \beta_j = \beta_{j|H_0}\ vs\ H_1: \beta_j \neq \beta_{j|H_0}$$

através da estatística de teste de Wald:

$$W = \frac{\hat{\beta}_j - \beta_{j|H_0}}{\hat{\sigma}_{\hat{\beta}_j}} \cap t_{(n-p)}, se\ H_0\ for\ verdadeira$$

sendo a região de rejeição, de acordo com a hipótese alternativa, dada (respetivamente) por:

$$|W_{\mathrm{OBS}}| > \mathrm{t}_{1-\frac{\alpha}{2};\mathrm{n-p}}$$

em que $\mathrm{t}_{1-\frac{\alpha}{2};\mathrm{n-p}}$ representa o quantil de ordem $1-\frac{\alpha}{2}$ da distribuição t de Student com $n-p$ graus de liberdade.

Testes à significância global do modelo construído serão dados pelas seguintes hipóteses:

$$H_0: \beta_2 = \beta_3 = \cdots = \beta_p = 0 \text{ vs } H_1: \exists j \in \{2,3,\ldots,p\} : \beta_j \neq 0$$

sendo a estatística de teste dada por:

$$F = \frac{MSR}{MSE} \cap F_{(p-1,n-\mathrm{p})}, se\ H_0\ for\ verdadeira$$

consoante as hipóteses em teste, sendo a região de rejeição em ambos os casos dada por

$$\mathrm{F_{OBS}} > \mathrm{F_{1-\alpha;(p-1,n-p)}}$$

em que $\mathrm{F_{1-\alpha;(p-1,n-p)}}$ é o quantil de ordem $1-\alpha$ da distribuição F com $p-1$ e $n-p$ graus de liberdade.

Podemos resumir os resultados deste teste de hipóteses através de uma tabela ANOVA como a seguinte (sendo o *p-value* dado por $\mathbb{P}(F > F_{obs})$):

Tabela 5.1 - Tabela ANOVA para o modelo clássico de regressão linear

Fonte de variação	Soma de quadrados	Graus de liberdade	Média de quadrados	Teste F
Regressão	$SSReg$	$p-1$	$MSReg = \frac{SSReg}{p-1}$	$F_{obs} = \frac{MSReg}{MSE}$
Resíduos	SSE	$n-p$	$MSE = \frac{SSE}{n-p}$	
Total	SST	$n-1$	$MST = \frac{SST}{n-1}$	

O modelo de regressão de Poisson

Noutras situações, podemos considerar que a população Y em estudo segue uma distribuição de Poisson, e modelar as variáveis aleatórias Y_i de acordo com uma regressão de Poisson. Este modelo possui muitas aplicações, sobretudo ao nível da análise de eventos envolvendo contagens, podendo também ser utilizado para modelar taxas de ocorrência de um dado evento no tempo, sendo tais taxas determinadas através de um quociente como Y/t.

Como

$$\mathbb{E}(\mathrm{Y_i}) = \mathbb{E}\left(\mathrm{Y_i}\middle|\vec{\mathrm{X}}_\mathrm{i} = \vec{\mathrm{x}}_\mathrm{i}\right) = \lambda_i > 0$$

deparamo-nos com um problema, uma vez que

$$\underbrace{\lambda_i}_{\text{Tem de assumir valores em } [0,+\infty[} = \underbrace{\beta_0 + \beta_1 \mathrm{X_{i1}} + \ldots + \beta_\mathrm{p} \mathrm{X_{ip}}}_{\text{Pode assumir qualquer valor em } \mathbb{R}}$$

problema este que pode ser resolvido através do recurso a uma transformação logarítmica:

$$\underbrace{ln(\lambda_i)}_{\text{Pode assumir qualquer valor em } \mathbb{R}} = \underbrace{\beta_0 + \beta_1 \mathrm{X_{i1}} + \ldots + \beta_\mathrm{p} \mathrm{X_{ip}}}_{\text{Pode assumir qualquer valor em } \mathbb{R}}$$

A função logarítmica tem em consideração a natureza multiplicativa de fenómenos registados em várias áreas do saber, pois:

$$\lambda_i = e^{\beta_0 + \beta_1 \mathrm{X_{i1}} + \ldots + \beta_\mathrm{p} \mathrm{X_{ip}}} = e^{\beta_0} e^{\beta_1 \mathrm{X_{i1}}} \ldots e^{\beta_\mathrm{p} \mathrm{X_{ip}}}$$

e, portanto, cada covariável X_{ij} exerce no valor esperado λ_i uma influência multiplicativa no valor de $e^{\beta_j X_{ij}}$. Adicionalmente, um aumento de uma unidade na da j-ésima covariável, X_{ij}, resulta numa influência multiplicativa dada por $e^{\beta_j} - 1$, uma vez que:

$$\left(e^{\beta_0} e^{\beta_1 X_{i1}} \dots e^{\beta_j (X_{ij}+1)} \dots e^{\beta_p X_{ip}}\right) - \left(e^{\beta_0} e^{\beta_1 X_{i1}} \dots e^{\beta_j X_{ij}} \dots e^{\beta_p X_{ip}}\right)$$
$$= \left(e^{\beta_j} - 1\right)\left(e^{\beta_0} e^{\beta_1 X_{i1}} \dots e^{\beta_j (X_{ij}+1)} \dots e^{\beta_p X_{ip}}\right)$$

De notar que, mesmo que as variáveis resposta Y_i sigam distribuições de Poisson (possivelmente distintas), se os seus valores médios λ_i forem elevados, estas distribuições poderão ser razoavelmente aproximadas pela distribuição Normal, e podemos recorrer a modelos de regressão linear já vistos, mais facilmente interpretáveis. Porém, como o fenómeno em estudo diz respeito a eventos cuja ocorrência é bastante rara (sinistros numa apólice e num dado ano civil), conclui-se que a esmagadora maioria das observações serão nulas, e assim sendo, teremos variáveis aleatórias associadas a contagens com valores médios diminutos, cenário no qual será recomendado o recurso a modelos de regressão de Poisson.

O modelo de regressão Gama

Comecemos por indicar que a função de densidade de probabilidade de uma variável aleatória Y com distribuição Gama com parâmetros $\alpha > 0$ e $\lambda > 0$ (ou seja, $Y \cap Gama\ (\alpha; \lambda)$) é dada por:

$$f_Y(y; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y}, y > 0$$

sendo $\Gamma(\alpha)$, a função gama, dada por

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}\, e^{-x} dx, \alpha \in \mathbb{R}^+$$

No modelo de regressão Gama, a função de ligação canónica é a função **inversa**, a qual se traduz em:

$$\frac{1}{\mu_i} = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}$$

Facilmente identificamos alertas aos quais devemos prestar atenção pois, simplificando este modelo de regressão para

$$\frac{1}{\mu_i} = \beta_0 + \beta_1 X_{i1} \Leftrightarrow \mu_i = \frac{1}{\beta_0 + \beta_1 X_{i1}}$$

facilmente se observa que esta função:

- Nem sempre é positiva, apesar de se ter $Y_i > 0, \forall i \in \{1,2,\dots,n\}$, pelo que temos de assegurar que $\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip} > 0, \forall i \in \{1,2,\dots,n\}$;
- Admite assíntotas verticais;

Em alternativa, podemos optar por uma função de ligação logarítmica, muito usada na prática:

$$ln(\mu_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

com uma interpretação similar à função de ligação vista em cima para o modelo de regressão de Poisson.

O modelo linear generalizado

Acabamos de ver modelos (como os de regressão de Poisson e de regressão Gama) nos quais os pressupostos do modelo clássico de regressão linear não são considerados aplicáveis. Na prática, verificam-se muitas situações onde a variável resposta Y, podendo até ser contínua, não satisfaz tais pressupostos, pois nestas:

- A variável resposta é não-negativa, ou seja, tem-se sempre $Y \geq 0$, tendo a distribuição de Y uma assimetria positiva/cauda à direita (e o suporte da distribuição Normal, a qual é simétrica, é $\mathbb{R}$)
- A variância de $Y_i|\vec{x}_i$ não é constante (por exemplo, em modelos de regressão de Poisson as variâncias coincidem com os valores médios, podendo então variar em conjunto);
- A relação entre o valor médio de $Y_i|\vec{x}_i$ e os valores observados das covariáveis $\vec{x}_i$ não é exclusivamente linear (nos parâmetros).

Num contexto atuarial isto costuma acontecer na análise:

- Da probabilidade de um dado sinistro ser considerado "extremo" ou não (regressão logística);
- Da frequência de sinistros num dado período de tempo (regressão de Poisson);
- De montantes de indemnizações de sinistros cobertos por seguros (regressão Gama).

Um olhar mais atento facilmente nos revela que tanto o modelo clássico de regressão linear como os modelos de regressão de Poisson e Gama partilham uma estrutura comum, como indicado na seguinte tabela:

Tabela 5.2 - Alguns exemplos de modelos lineares generalizados

Modelo de regressão	Distribuição de Y_i	Especificação
Linear (clássico)	$\text{Normal}(\mu_i, \sigma^2)$	$\mu_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$
Poisson	$\text{Poisson}(\lambda_i)$	$ln(\lambda_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$
Gama	$\text{Gama}(\alpha, \lambda_i)$	$\frac{1}{\mu_i} = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$

assumindo-se em todos os casos a independência das variáveis aleatórias Y_i. Note-se que:

- Todas as distribuições em cima mencionadas pertencem à Família Exponencial;

- Em todos os modelos em cima apresentados, uma transformação do valor esperado é modelada através de uma combinação linear de covariáveis.

Podemos então, e após esta introdução, apresentar o conceito de modelo linear generalizado. Um modelo linear generalizado resulta da adequada combinação dos três componentes que dele fazem parte:

- A componente aleatória, a qual especifica a distribuição condicional da variável resposta $Y_i|\vec{x}_i$;
    - Na formulação original de Nelder & Wedderburn, a distribuição de cada variável aleatória Y_i ($i \in \{1,2,\ldots,n\}$) pertence à família exponencial de distribuições (e daí a relevância desta família);
- A componente sistemática ou preditor linear $\eta_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$, enquanto função linear das covariáveis;
    - Os regressores X_{ij} podem ser funções das variáveis explicativas, permitindo (a título de exemplo) a existência de regressores polinomiais, variáveis *dummy*, interações, etc.;
    - Uma das vantagens dos modelos lineares generalizados é o facto de, apesar de serem de aplicação mais vasta, contemplarem um preditor linear com natureza similar à de um modelo clássico de regressão linear;
- A função de ligação (*link function*) $g(.)$, uma função real de variável real lisa (*smooth*) e invertível, a qual possui este nome porque liga as duas componentes anteriores, transformando o valor esperado da variável resposta, $\mu_i = \mathbb{E}(Y_i)$ no preditor linear, η_i:

$$g(\mu_i) = \eta_i$$

Sendo $g(.)$ invertível, podemos naturalmente escrever a função média (*mean function*) como sendo

$$g(\mu_i) = \eta_i \iff \mu_i = g^{-1}(\eta_i)$$

Porém, e como já visto, o modelo associado a μ_i será (quase sempre) mais complexo do que o modelo associado a η_i, o qual corresponde a uma combinação linear dos regressores.

Existe, para cada distribuição da família exponencial/modelo linear generalizado em estudo, uma função de ligação mais "natural" e que possibilita o acesso a propriedades estatísticas desejáveis, sendo esta a função de ligação canónica. De uma forma rigorosa, a função de ligação canónica é a função de ligação que faz com que $\eta \equiv \theta$. Porém, e apesar das vantagens associadas a funções de ligação canónicas, nada nos impede de escolhermos outras funções de ligação, sobretudo se tal escolha nos parecer mais justificada.

De notar que, sendo $g(.)$ uma função real de variável real e Y uma variável aleatória, nada obriga a que $\mathbb{E}\big(g(Y)\big) = g\big(\mathbb{E}(Y)\big)$, pelo que $\mathbb{E}\big(g(Y_i|\vec{x}_i)\big) \neq g\big(\mathbb{E}(Y_i|\vec{x}_i)\big)$. Assim, se num dado contexto não forem satisfeitos os pressupostos do modelo clássico de regressão linear, temos duas soluções possíveis:

- Aplicar uma transformação aos dados, isto é, à variável resposta, Y, de modo a se (tentar) satisfazer tais pressupostos;
    - Esta opção corresponde a um modelo clássico de regressão linear com dados transformados;
    - Por exemplo: $\mathbb{E}\big(ln(Y)\big) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$;
- Aplicar uma transformação ao modelo, isto é, à expressão (inicialmente linear) a aplicar aos dados;

- Esta opção corresponde a um modelo linear generalizado;
- Por exemplo: $ln\big(\mathbb{E}(Y)\big) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$.

Modelos lineares generalizados trazem consigo novas possibilidades, pois:

- A variável resposta Y pode seguir outras distribuições[9] que não a Normal;
- O valor esperado condicional $\mathbb{E}(Y_i|\vec{x}_i)$ pode corresponder a uma função real de variável real $g^{-1}(.)$ (obviamente invertível) da combinação linear $\eta_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$, não tendo necessariamente de ser igual a esta combinação linear;
- Sendo a distribuição de Y pertencente à família exponencial de distribuições, deixa de ser imposta a condição da homoscedasticidade[10] pois, como já visto, nesta família, a identidade $V(\mu_i) = b''(\theta_i)a(\varphi_i)$ indica-nos que diferentes valores para as covariáveis $\vec{x}_i$ podem fazer mudar o valor médio μ_i, o que por sua vez pode causar alterações na variância $V(\mu_i)$, ou seja, modelos lineares generalizados também modelam a variância enquanto função das variáveis explicativas e/ou preditivas.

Assim sendo, quando pretendermos ajustar modelos de regressão a dados e os pressupostos do modelo clássico de regressão linear são infringidos, podemos cogitar o uso de modelos lineares generalizados.

Porém, modelos lineares generalizados trazem também novas preocupações. O ajustamento de modelos desta natureza através do método de máxima verosimilhança a conjuntos de dados observados envolve muitas vezes a resolução de equações não lineares, o que exigirá a aplicação de métodos numéricos, isto é, o uso de algoritmos que permitam a obtenção de soluções aproximadas para estas equações, devendo o erro decorrente destas aproximações ser diminuto. Para além disso, o diagnóstico de modelos lineares generalizados não é tão direto e simples como o diagnóstico de modelos clássicos de regressão linear. Por último, existem propriedades muito interessantes no modelo clássico de regressão linear que deixam de ser válidas em modelos lineares generalizados, como, por exemplo, a independência entre os coeficientes de regressão e da variância dos resíduos.

Parâmetros de um MLG - estimação e propriedades

Modelos lineares generalizados são ajustados a conjuntos de dados através do método de máxima verosimilhança, o qual fornece não só estimativas dos coeficientes de regressão, mas também estimativas assintóticas (ou seja, para amostras grandes) dos erros-padrão (amostrais) associados aos mesmos. A log-verosimilhança associada a uma variável aleatória Y (com valor observado y) pertencente à família exponencial de distribuições é dada por:

[9] Apesar de ser possível o recurso a transformações de dados que nos aproximem da Normalidade (como a transformação logarítmica que será ultilizada na prática, ou um membro da família de transformações de Box-Cox), ou à teoria que garante a normalidade assintótica ds estimadores de mínimos quadrados dos coeficientes em condições bastantes gerais (Srivastava). Também nos modelos lineares generalizados se dá a existência de resultados assintóticos, sendo contudo estes menos robustos (pois estas aproximações são muito mais sensíveis a afastamentos ou desvios à distribuição postulada).

[10] Também é possível, no modelo de regressão linear, o recurso a transformações que estabilizem a variância dos resíduos.

$$l_i(\vec{\beta}) = ln\big(\mathcal{L}(y;\theta_i,\phi_i)\big) = \frac{y\theta_i - b(\theta_i)}{a(\phi_i)} + c(y,\phi_i)$$

e a log-verosimilhança associada a uma amostra aleatória $\vec{\mathbf{Y}} = (Y_1, Y_2, \ldots, Y_n)$ (com valores observados $\vec{\mathbf{y}} = (y_1, y_2, \ldots, y_n)$) pertencente à família exponencial de distribuições é dada por:

$$l(\vec{\beta}) = ln\left(\mathcal{L}(\vec{\mathbf{y}};\vec{\boldsymbol{\theta}},\vec{\boldsymbol{\phi}})\right) = \sum_{i=1}^{n}\left(\frac{y\theta_i - b(\theta_i)}{a(\phi_i)} + c(y,\phi_i)\right) = \sum_{i=1}^{n} l_i(\vec{\beta})$$

sendo $\vec{\boldsymbol{\theta}} = (\theta_1, \theta_2, \ldots, \theta_n)$ e $\vec{\boldsymbol{\phi}} = (\phi_1, \phi_2, \ldots, \phi_n)$.

Um modelo linear generalizado com função de ligação $g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$ expressa, portanto, n variáveis aleatórias Y_i com valores observados y_i em função de um número p (bem) menor de regressores.

O nosso problema passa então a ser dado por

$$\max_{\vec{\beta}} l(\vec{\beta})$$

Este é um problema de otimização que não pode ser resolvido analiticamente, sendo então necessária a aplicação de algoritmos que permitam uma resolução numérica, uma vez que resulta nas seguintes condições de 1ª ordem:

$$\nabla l(\vec{\beta}) = \vec{\mathbf{0}} \Leftrightarrow \frac{\partial l(\vec{\beta})}{\partial \beta_j} = 0 \Leftrightarrow \sum_{i=1}^{n} \frac{\partial l_i(\vec{\beta})}{\partial \beta_j} = 0 \ (\forall\, j \in \{1,2,\ldots,p\})$$

Iremos então descrever de seguida, e de forma algo geral, de que forma este problema tende a ser resolvido.

Definição: *Ao gradiente da função de log-verosimilhança (isto é, ao vetor composto pelas derivadas parciais de 1ª. ordem da função de log-verosimilhança em ordem a cada um dos parâmetros pertencentes ao vetor $\vec{\boldsymbol{\theta}} = (\theta_1, \theta_2, \ldots, \theta_n)$) dá-se o nome de função score de Fisher, sendo esta função denotada por:*

$$\vec{\boldsymbol{s}}(\vec{\boldsymbol{\theta}}) = \nabla ln\left(\mathcal{L}(\vec{\mathbf{y}};\vec{\boldsymbol{\theta}})\right) = \frac{\partial ln\left(\mathcal{L}(\vec{\mathbf{y}};\vec{\boldsymbol{\theta}})\right)}{\partial \vec{\boldsymbol{\theta}}}$$

De notar que, sendo o gradiente um vetor, a função score será naturalmente uma função vetorial.

Verifica-se que se a log-verosimilhança for uma função côncava, para encontrar os estimadores de máxima verosimilhança basta encontrar uma maneira de resolver o sistema dado por:

$$\vec{\boldsymbol{s}}\left(\vec{\widehat{\boldsymbol{\theta}}}\right) = \vec{\mathbf{0}}$$

(quase sempre não-linear). É este o nosso problema atual – resolver $\nabla l(\vec{\beta}) = \vec{\mathbf{0}}$.

Definição: *À matriz de covariâncias da função score dá-se o nome de matriz de informação de Fisher, a qual é denotada por*

$$\boldsymbol{I}(\vec{\boldsymbol{\theta}}) = \text{var}\left(\vec{\boldsymbol{s}}(\vec{\boldsymbol{\theta}})\right)$$

Sob algumas condições de regularidade, verifica-se que

$$\boldsymbol{I}(\vec{\boldsymbol{\theta}}) = \left[-\mathbb{E}\left(\frac{\partial^2 \ln\left(\mathcal{L}(\vec{\boldsymbol{\theta}})\right)}{\partial\theta_i \partial\theta_j}\right)\right]_{i,j\in\{1,2,\ldots,n\}}$$

Se quisermos resolver numericamente o sistema $\vec{\boldsymbol{s}}\left(\vec{\widehat{\boldsymbol{\theta}}}\right) = \vec{\mathbf{0}}$, podemos recorrer ao método de Newton-Raphson, no qual se considera a aproximação

$$\vec{\boldsymbol{s}}(\vec{\boldsymbol{\theta}}) = \vec{\boldsymbol{s}}(\vec{\boldsymbol{\theta}}_0) + \frac{\partial\vec{\boldsymbol{s}}(\vec{\boldsymbol{\theta}})}{\partial\vec{\boldsymbol{\theta}}}(\vec{\boldsymbol{\theta}} - \vec{\boldsymbol{\theta}}_0)$$

como base para o processo iterativo dado por

$$\vec{\boldsymbol{\theta}}_k = \vec{\boldsymbol{\theta}}_{k-1} - \mathbf{H}^{-1}(\vec{\boldsymbol{\theta}}_{k-1})\vec{\boldsymbol{s}}(\vec{\boldsymbol{\theta}}_{k-1})$$

pois note-se que

$$\frac{\partial\vec{\boldsymbol{s}}(\vec{\boldsymbol{\theta}})}{\partial\vec{\boldsymbol{\theta}}} = \mathbf{H}(\vec{\boldsymbol{\theta}})$$

isto é, a derivada da função *score*, já de si um gradiente, corresponde à matriz Hessiana da função de log-verosimilhança.

Uma outra alternativa, inicialmente proposta por Fisher, passa por substituir a matriz Hessiana pelo seu valor esperado (a matriz de informação), obtendo-se então o método de *scoring* de Fisher, no qual as iterações são dadas por

$$\vec{\boldsymbol{\theta}}_k \leftarrow \vec{\boldsymbol{\theta}}_{k-1} + \mathbf{I}^{-1}(\vec{\boldsymbol{\theta}}_{k-1})\vec{\boldsymbol{s}}(\vec{\boldsymbol{\theta}}_{k-1})$$

No contexto do nosso problema, o processo iterativo a considerar será definido mediante

$$\vec{\boldsymbol{\beta}}_k \leftarrow \vec{\boldsymbol{\beta}}_{k-1} + \mathbf{I}^{-1}(\vec{\boldsymbol{\beta}}_{k-1})\vec{\boldsymbol{s}}(\vec{\boldsymbol{\beta}}_{k-1})$$

e estas iterações deverão, após a introdução de estimativas iniciais adequadas, ser repetidas até os coeficientes (estimados) da regressão estabilizem, obtendo-se assim convergência para os estimadores de máxima verosimilhança de $\vec{\boldsymbol{\beta}}$.

O método de *scoring* de Fisher corresponde:

- Ao algoritmo IWLS - *Inverse-Weighed Least Squares*;
- Ao método Newton-Raphson (em funções de ligação canónicas).

Em relação ao parâmetro ϕ, este pode ser estimado pelo método dos momentos através de:

$$\hat{\phi} = \frac{1}{n - p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\mathbb{V}(\hat{\mu}_i)}$$

Testes de hipóteses, intervalos de confiança e inferências

Como acabamos de ver, modelos lineares generalizados são ajustados a conjuntos de dados pelo método da máxima verosimilhança. É possível demonstrar que os estimadores de máxima verosimilhança para $\vec{\boldsymbol{\beta}}$, $\overrightarrow{\widehat{\boldsymbol{\beta}}}$, gozam das seguintes propriedades, consideradas ótimas:

- São estimadores assintoticamente centrados: $\mathbb{E}\left(\overrightarrow{\widehat{\boldsymbol{\beta}}_n}\right) \to \vec{\boldsymbol{\beta}}$, quando $n \to +\infty$;
- São estimadores consistentes: $\overrightarrow{\widehat{\boldsymbol{\beta}}_n} \to \vec{\boldsymbol{\beta}}$, em probabilidade, quando $n \to +\infty$;
- São assintoticamente Normais ou Gaussianos: $\overrightarrow{\widehat{\boldsymbol{\beta}}_n} \cap \mathcal{N}_p\left(\vec{\boldsymbol{\beta}}, \mathbf{I^{-1}}\left(\vec{\boldsymbol{\beta}}\right)\right)$;
- São assintoticamente eficientes, face a outros estimadores.

Porém, por mais interessantes que sejam, os estimadores de máxima verosimilhança não são suficientes para analisar de forma aprofundada um dado modelo linear generalizado. Como seria de esperar, estimativas pontuais só são verdadeiramente informativas se o erro[11] associado às mesmas for conhecida. Isto é particularmente relevante para os erros padrão dos coeficientes, isto é, os desvios-padrão das suas estimativas. Com estes dois elementos (estimativas e medidas de erro) torna-se possível a realização de testes de hipóteses e a construção de intervalos de confiança.

Entretanto, atente-se no facto de que, ao contrário do que acontece no modelo clássico de regressão linear, onde $cov\left(\vec{\widehat{\boldsymbol{\beta}}}\right) = \mathbf{I^{-1}}\left(\vec{\boldsymbol{\beta}}\right) = \sigma^2\left(\boldsymbol{X^T X}\right)^{-1}$, em modelos lineares generalizados, as (co)variâncias dos estimadores dos coeficientes dependem dos próprios valores destes coeficientes, o que não é uma situação propriamente ideal, pois é precisamente por estes valores serem desconhecidos que estão a ser estimados!

Para contornar este problema podemos determinar a matriz de informação de Fisher em $\vec{\boldsymbol{\beta}} = \overrightarrow{\widehat{\boldsymbol{\beta}}}$, substituindo os coeficientes desconhecidos pelos seus valores estimados. Nesta eventualidade, temos:

$$\mathbf{I}\left(\vec{\boldsymbol{\beta}}\right) = -\mathbf{H}\left(\vec{\boldsymbol{\beta}}\right) = \boldsymbol{X^T V X}$$

$$\boldsymbol{V} = \begin{bmatrix} V_1 & 0 & \vdots & 0 \\ 0 & V_2 & \vdots & 0 \\ \dots & \dots & \ddots & \dots \\ 0 & 0 & \vdots & V_n \end{bmatrix}$$

$$V_i = var(Y_i)$$

Como no modelo clássico de regressão linear $var(Y_i) = \sigma^2, \forall i \in \{1,2,\dots,n\}$, chegamos ao resultado em cima descrito, facilmente comprovamos que o pressuposto da homoscedasticidade traz uma maior simplicidade, dado que neste caso a precisão dos estimadores de mínimos quadrados não depende dos próprios valores destes coeficientes desconhecidos.

[11] Ou uma sua estimativa.

A substituição $\vec{\boldsymbol{\beta}} = \vec{\widehat{\boldsymbol{\beta}}}$ resulta em $\boldsymbol{V} = \widehat{\boldsymbol{V}}$. Sendo v_{jj} o j-ésimo elemento da matriz $\left(\boldsymbol{X^T V X}\right)^{-1}$, ter-se-á que $\hat{\sigma}_{\hat{\beta}_j} = \sqrt{v_{jj}}$ e, assim sendo, podemos apresentar a estatística de Wald como sendo

$$W = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \cap \mathcal{N}(0,1)$$

Alternativamente, podemos observar que $W^2 \cap \chi_1$, sendo esta relação assintótica em ambos os casos. Assim sendo, podemos recorrer mais uma vez ao teste de Wald para testar as hipóteses

$$H_0: \beta_j = \beta_{j|H_0} \; vs \; H_1: \beta_j \neq \beta_{j|H_0}$$

$$W = \frac{\hat{\beta}_j - \beta_{j|H_0}}{\hat{\sigma}_{\hat{\beta}_j}} \cap \mathcal{N}(0,1), se \; H_0 \; for \; verdadeira$$

Concluímos que a j-ésima covariável ($j \in \{1,2,\dots,p\}$) é estatisticamente significativa para o modelo se o valor observado da estatística de teste pertencer à região de rejeição, isto é, se:

$$|W_{\text{OBS}}| > \text{z}_{1-\frac{\alpha}{2}}$$

Podemos também considerar os seguintes intervalos a $(1-\alpha) \times 100\%$ de confiança para β_j, e extrair conclusões idênticas se o valor 0 **não** pertencer a este intervalo:

$$\beta_j \in \left[\hat{\beta}_j \mp \text{z}_{1-\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{\beta}_j}\right]$$

Contudo, podemos pretender construir intervalos a $(1-\alpha) \times 100\%$ de confiança **não** para β_j, mas sim para uma sua transformação como e^{β_j} (do nosso interesse quando fazemos uso de funções de ligação logarítmicas, como será o caso). Sendo esta uma transformação crescente, podemos escrever:

$$e^{\beta_j} \in \left[e^{\hat{\beta}_j - \text{z}_{1-\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{\beta}_j}}, e^{\hat{\beta}_j + \text{z}_{1-\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{\beta}_j}}\right]$$

Para verificar se as diferenças entre um modelo inicial generalizado (com p parâmetros) e o sub-modelo a partir deste construído (com q parâmetros) são estatisticamente significativas, podemos considerar uma estatística mais popular do que a estatística de Wald, em modelos encaixados – a estatística de razão de verosimilhanças. Esta é dada por:

$$\lambda = -2 \ln\left(\frac{\max\limits_{H_0} \mathcal{L}(\vec{\boldsymbol{\beta}})}{\max\limits_{H_0 \cup H_1} \mathcal{L}(\vec{\boldsymbol{\beta}})}\right)$$

Sendo S o conjunto de índices dos parâmetros presentes no modelo global, mas não no sub-modelo, esta será a estatística associada ao seguinte teste de hipóteses:

$$H_0: \beta_j = 0, \forall j \in S \; vs \; H_1: \exists j \in S: \beta_j \neq 0$$

Pelo teorema de Wilks, esta estatística, sob a validade de H_0, segue uma distribuição χ^2 com $p-q$ graus de liberdade (assintoticamente). A região de rejeição será naturalmente dada por $|\lambda| > \chi^2_{1-\alpha,\text{p}-\text{q}}$.

Na realização de inferências em modelos lineares generalizados é comum a consideração de dois modelos de natureza particular que, embora demasiado extremos para serem implementados na prática, são úteis para a realização de comparações. Estes modelos são:

- O modelo nulo, no qual o valor ajustado a cada indivíduo ou unidade estatística é o mesmo, não havendo lugar a covariáveis;
- O modelo completo, no qual há lugar a n covariáveis e, portanto, a n parâmetros, um para cada observação.

Podemos então definir o desvio ou *deviance* de um dado modelo linear generalizado como sendo

$$\mathcal{D} = -2\ln\left(\frac{\mathcal{L}(\text{modelo em estudo})}{\mathcal{L}(\text{modelo completo})}\right)$$

Em particular:

- O desvio nulo é denominado, no R, de *null deviance* e corresponde ao desvio do modelo nulo;
- O desvio do modelo proposto/ajustado é denominado, no R, de *residual deviance*.

Modelos lineares generalizados não compreendem um coeficiente de determinação (R^2) que seja de caráter universal. Porém, podemos recorrer a um indicador análogo, que também mede a diferença entre a capcidade explicativa e preditiva do modelo considerado face ao modelo nulo. Defina-se então o *pseudo-*R^2 de McFadden como sendo:

$$R^2_{MLG} = 1 - \frac{\mathcal{D}_{modelo\ construído}}{\mathcal{D}_{modelo\ nulo}}$$

Nota: Temáticas como a o diagnóstico e a validação de modelos lineares generalizados, bem como a avaliação de medidas de erro e eventual escolha de modelos, será apresentada de uma forma prática/aplicada na secção seguinte.

Modelação tarifária em Atuariado

Introdução à modelação atuarial. Modelos tarifários

O Atuariado distingue-se da Estatística pois tem em maior consideração o seu contexto de aplicação – na esmagadora maioria dos casos, este é o contexto dos seguros. No entanto, os modelos atuariais têm sobretudo uma natureza estatística pois são, regra geral, estocásticos.

Um *modelo atuarial*, enquanto modelo matemático, corresponde a uma descrição simplificada da realidade. Modelos atuariais são construídos e utilizados por atuários para formar uma opinião fundamentada e recomendar planos de ação para lidar com eventos futuros e aleatórios.

Existe na modelação atuarial um especial foco na definição de prémios, ou seja, na tarifação de seguros, pois a principal preocupação de uma seguradora é a de avaliar de forma correta e adequada o preço de transferência de cada risco que decide aceitar, isto é, o custo do risco - prémio puro.

Tais modelos tarifários surgem graças à existência de conjuntos de dados históricos de qualidade, beneficiando ainda de uma definição clara do posicionamento/estratégia do produto, do desenho de coberturas e de análises técnicas adicionais, devendo resultar na obtenção de prémios suficientes para cobrir as responsabilidades da seguradora com elevada probabilidade, e ainda encargos adicionais (por exemplo, administrativos, de distribuição, de custo de capital).

Podemos conjugar as preocupações tradicionalmente associadas à construção de modelos atuariais com as preocupações típicas de projetos de *Data Science* e obter o seguinte processo (de natureza iterativa) de modelação quantitativa:

1. Passos prévios:
    a. Geração/observação, registo e armazenamento de dados;
    b. Transferência/obtenção dos dados;
    c. Compreensão da natureza dos dados.
2. Pré-processamento dos dados:
    a. Análise exploratória de dados (inicial, para detetar erros, omissões e inconsistências);
    b. Limpeza de dados;
    c. Integração de dados;
    d. Redução de dados;
    e. Transformação de dados.
3. Processamento/análise dos dados:
    a. Análise exploratória de dados (para confirmar o correto pré-processamento dos dados, bem como descobrir possíveis relações/informações de interesse sobre os mesmos);
    b. Identificação de modelos apropriados aos dados;
    c. Ajustamento de modelos apropriados aos dados;
    d. Validação de modelos construídos;
    e. Seleção e modificação de modelos;
    f. Interpretação do modelo obtido;
    g. Geração de estimativas/previsões relevantes.

4. Pós-processamento dos dados:
    a. Extração de conclusões;
    b. Apresentação de conclusões;
    c. Consideração de aspetos de natureza comercial/empresarial e estudo de impactos (adequação de prémios a cada segmento-alvo, definição de *loadings* e aplicação de princípios de cálculo do prémio);
    d. Possível implementação do modelo.

Dados são, portanto, importantes (desde que apropriados) em Atuariado, uma vez que são registos que permitem, mediante análises e interpretações sensatas, obter informações - conhecimentos que podem servir como base para ações futuras bem-sucedidas.

Na construção de modelos tarifários, as perdas totais são representadas pela variável aleatória S, a qual por sua vez pode ser vista como o produto de duas outras variáveis aleatórias – N, a frequência de sinistros (número de perdas ocorridas num dado período de tempo) numa dada apólice, e X, a severidade de um sinistro (o montante de uma dada perda ocorrida no período de tempo em estudo).

Pré-processamento de dados

Introdução. Importação de dados

Neste projeto foi feito uso da linguagem R e, em particular, do *Integrated Development Environment* (IDE) RStudio, o qual é provavelmente o mais popular no universo de utilizadores do R, e garantidamente dentro da ASP. Esta revelou-se uma boa escolha, uma vez que o RStudio possui ferramentas adicionais interessantes, não presentes no R base, duas das quais foram utilizadas neste projeto - o RMarkdown e os R Projects.

Antes de explicar os benefícios extraídos de cada uma destas ferramentas, importa descrever as nossas pretensões neste projeto, em termos de estrutura do mesmo, destacando a transparência e a eficiência como princípios fundamentais a seguir. Uma preocupação fundamental é a de realizar tarefas apenas uma vez, e da forma mais adequada possível. Pela relevância deste assunto, a seguinte explicação está no texto principal e não nos anexos.

Sendo o trabalho de alguém com formação quantitativa a realização de análises em diferentes contextos, facilmente se chega à conclusão de que é boa prática manter tais projetos separados, atribuindo-lhes pastas ou diretórios específicos, embora alojados num sítio comum. Esta prática confere organização ao nosso trabalho, o qual é não raras vezes complexo.

Os R Projects permitem associar a cada projeto/análise um diretório específico, como desejado. Adicionalmente, esta ferramenta atribui a cada projeto um ficheiro com extensão .RProj que, quando alvo de duplo clique, permite abrir o projeto em questão numa instância isolada e "limpa" do R, de maneira a que a análise atual não seja afetada pelos resultados de outras análises, levadas a cabo num passado recente no RStudio. Por último, ao fazer uso dos R Projects (sempre através do ficheiro .RProj, como recomendado), o RStudio define o diretório de referência original como sendo o do projeto, de forma automática e sem

necessidade de ajustes manuais, o que facilita a execução do projeto noutos computadores e, portanto, a partilha do mesmo.

No entanto, e para além de organizado, o nosso trabalho deve também ser replicável. Ao ser cada vez mais central, sobretudo em investigações científicas, este assunto foi também alvo da nossa atenção neste trabalho, o qual introduziu esta temática na ASP, com as restrições de confidencialidade aplicáveis (ocorrendo a partilha de projetos apenas no interior da ASP). Para saber se uma dada análise é replicável, o autor da mesma deve procurar responder à seguinte pergunta:

"*Se disponibilizar a outra pessoa os conjuntos de dados alvo de análise, o código por mim escrito e os resultados deste código, bem como as interpretações destes resultados, será que quem receberá tais ficheiros conseguirá reproduzir os passos por mim seguidos e obter resultados iguais ou, pelo menos, bastante similares, no seu computador?"*

Esta preocupação com a replicabilidade também permite ao autor de uma análise perceber no futuro os passos que seguiu aquando da elaboração da mesma, conferindo-lhe assim uma maior qualidade (uma vez que pode mais facilmente ser alvo de verificações por outros colaboradores numa organização, por entidades de supervisão e regulação, ou por entidades responsáveis por auditorias).

A replicabilidade de um projeto é assegurada pelos R Projects (os quais facilitam a partilha de análises) e, também, pelo RMarkdown. O Markdown é uma linguagem de *mark-up* que permite aplicar a texto comum (*plain text*) formatação básica (como, por exemplo, títulos, negrito, listas numeradas). Por sua vez, o RMarkdown é a implementação do Markdown no R, estando presente e integrada no RStudio (mas não no R base).

O recurso ao RMarkdown permite a criação de documentos com extensão .Rmd que intercalem os três elementos em cima mencionados: código em R (ou outras linguagens suportadas), *outputs* associados a este código e texto em Markdown (o qual permite descrever os raciocínios que conduziram a tais *outputs*). O resultado é a obtenção de *notebooks* que possuem também caráter explicativo, em detrimento de *scripts* que servem apenas para executar código. Adicionalmente, a missão de profissionais como atuários, estatísticos, cientistas de dados e profissionais de áreas afins é o de efetuar análises de dados, e não apenas o de escrever código. Por outras palavras, pode haver lugar ao desenvolvimento de *software*, mas este é apenas um meio para alcançar o que se pretende: atribuir um significado a dados em análise, e sugerir possíveis ações futuras.

Assim sendo, é sem surpresas que se indica que este projeto foi conduzido internamente na ASP através do recurso ao RMarkdown e dos R Projects. Em particular, o RMarkdown permite compilar (*knit*), isto é, exportar conteúdos presentes em ficheiros .Rmd para outros formatos mais populares, como HTML, PDF, DOC (MS Word) e PPT (MS PowerPoint), tendo-se obtido desta forma ficheiros DOC que não foram originalmente criados no MS Word, mas cujos conteúdos foram diretamente relevantes para este documento final, este sim alvo de edição no MS Word.

É ainda possível, através de algumas utilidades[12], criar *scripts* contendo o código presente em ficheiros RMarkdown. Assim sendo, o RMarkdown tem a vantagem de ser uma solução completa que serve, mediante exportação de conteúdos, de "ponto médio" entre relatórios (apropriados para apresentação de resultados a outros *stakeholders*) e *scripts* (adequados para execução de código).

[12] Como a disponível em https://bookdown.org/yihui/rmarkdown-cookbook/purl.html

Houve também neste trabalho a preocupação de evitar o recurso a *packages*, os quais requerem não raras vezes instalação e trazem consigo sintaxe própria e, por isso, alguma incoerência no estilo do código criado. Tal estratégia levou, invariavelmente, à criação de algumas funções já implementadas em outros pacotes.

O princípio DRY (*Don't Repeat Yourself*) é um dos conceitos mais importantes em programação, sendo também relevante para este trabalho pois, apesar da essência do mesmo não ser de natureza computacional, a execução de código é essencial para cumprir os objetivos do projeto.

Foram então seguidas algumas regras práticas que permitem perseguir o objetivo de abstrair procedimentos, nomeadamente,

- Não repetir comandos idênticos, antes usar ciclos (havendo quem, no R, recomende antes o uso de uma função da família **apply**);
- Não repetir ciclos idênticos, antes criar uma função;
- Não definir funções à medida que precisamos delas, antes armazená-las num *script* ou módulo a importar no início de cada fase da nossa análise.

Para além destas, também são recomendadas outras práticas, tais como:

- Definir no início do *notebook* em **RMarkdown** objetos gerais (como constantes) que não tenham sido armazenados em *scripts*;
- Fixar a *seed* responsável pelo funcionamento dos geradores de números aleatórios.

Por outro lado, não devemos exagerar nesta abstração. Por exemplo, uma função deve ter um propósito específico ou, pelo menos, não muito amplo; tarefas de natureza distinta devem, se necessário, ser alvo de funções distintas.

O elevado número de funções criadas encontra então justificação na utilidade que as mesmas trazem para generalizar operações e acelerar a execução de tarefas, permitindo assim poupanças no esforço a dispender para fazer face ao trabalho com o qual nos deparamos.

Depois de criado, um *script* deve ser armazenado. Mas onde? O diretório deste projeto de tarifação representa toda uma hierarquia de subdiretórios, hierarquia essa que pode ser vista consultando o anexo relevante, de seu nome **Hierarquia e estrutura do projeto de tarifação**.

Tendo em conta as vantagens associadas ao RMarkdown e aos R Projects, bem como a hierarquia de pastas visíveis em anexo, é fácil entender os motivos que nos levaram a conduzir a nossa análise de uma forma sequencial, através de três ficheiros em RMarkdown, através dos quais foram gerados tanto ficheiros HTML como documentos editáveis no MS Word. Estes ficheiros são:

1. **01_pre_processamento.Rmd** (em **./pipeline/1_dados**), destinado ao pré-processamento de dados;
2. **02_analise_exploratoria_dados.Rmd** (em **./pipeline/2_aed**), destinado à análise exploratória de dados;
3. **03_modelacao.Rmd** (em **./pipeline/3_modelos**), destinado à construção de modelos atuariais que permitam a obtenção de uma estrutura tarifária.

Apesar do processo de Ciência de Dados anteriormente detalhado ser iterativo, a execução destes ficheiros deve ser feita por esta ordem.

Limpeza, integração, redução e transformação de dados

Importados os dados originais, chegamos à fase de **pré-processamento**, tema que compreende sobretudo quatro passos essenciais:

- A **limpeza de dados**, na qual, por exemplo, se corrigem dados registados inicialmente com erros;
- A **integração de dados**, na qual se procura efetuar a agregação de dados oriundos de várias fontes num "registo unificado";
- A **redução de dados**, na qual se eliminam, por exemplo, registos duplicados (isto também corresponde à limpeza de dados);
- A **transformação de dados**, na qual, por exemplo, se normalizam ou *standardizam* os dados (isto é, se reduzem os dados a valores em $[-1,1]$ ou $[0,1]$, ou a valores com média nula e desvio-padrão unitário).

Apesar destes passos estarem indicados de forma ordenada, o que se verifica muitas vezes (em projetos de Ciência de Dados) é que os mesmos se sucedem de forma **iterativa**.

Para pré-processarmos os dados, precisamos inevitavelmente de ter uma ideia do aspecto dos mesmos. Por isso, podemos começar por analisar informações inerentes ao próprio conjunto de dados em si, através da aplicação dos comandos **head**(), **dim**(), **names**(), **class**() e **str**() tanto a `carteira` como a `sinistros`. No geral, os passos, processos e raciocínios seguidos podem ser consultados nos diversos anexos disponíveis (**Tarefas adicionais de pré-processamento**, **Determinação da duração da exposição ao risco**, **Ainda sobre o pré-processamento**, **Integração de dados**, **Criação de conjuntos de treino, validação e teste**, e **Exportação de dados pré-processados**).

Nota: O conteúdo disposto em **ddcc** diz respeito a descrições de localizações geográficas a serem integrados na nossa análise, pelo que não serão analisados da mesma forma que **carteira** e **sinistros**.

Introdução ao processamento de dados

Efetuado o pré-processamento dos dados na sua totalidade, passamos agora à fase de **processamento**, isto é, de **análise** dos dados. Nesta fase, iremos englobar:

- A **análise exploratória de dados**;
- A obtenção de **estimativas**, bem como a realização de **inferências** e de **testes de hipóteses** (hipóteses essas que podem ter sido levantadas na fase de análise exploratória de dados);
- A construção de **modelos estatísticos**, que permitam modelar variáveis resposta através de variáveis explicativas ou preditivas (possivelmente, modelos lineares generalizados);
- A realização de possíveis **comparações entre modelos** (ao nível dos seus desempenhos, em conjuntos de validação, e só se forem construídos vários modelos);
- A implementação de um ***modelo final***.

A análise exploratória de dados (abreviadamente, AED, ou EDA, de *Exploratory Data Analysis*, em inglês) é uma fase crítica na análise de conjuntos de dados, pois compreende todos os métodos estatísticos **não**-formais (leia-se, modelação e realização de inferências) através dos quais podemos "olhar" para os dados, nomeadamente:

- A detecção de erros;
- A detecção de (possíveis) relações entre variáveis (de natureza aleatória), sobretudo ao nível da "direcção" e da "intensidade" das mesmas;
- A escolha preliminar de (tipos de) modelos.

Como diz Hadley Wickham no seu manual *online*, *R for Data Science*,

"EDA is not a formal process with a strict set of rules. More than anything, EDA is a state of mind. During the initial phases of EDA you should feel free to investigate every idea that occurs to you. Some of these ideas will pan out, and some will be dead ends. As your exploration continues, you will home in on a few particularly productive areas that you will eventually write up and communicate to others."

É neste espírito que as análises exploratórias foram realizadas. Estas podem ser vistas nos anexos relevantes (secção **Análise exploratória de dados**).

Modelação da frequência de sinistros

Alguns passos adicionais

Vamos agora preocupar-nos com uma especificidade da modelação atuarial, a de construção de categorias para cada variável explicativa/preditiva (mesmo que as mesmas sejam contínuas, levando tal procedimento a uma perda de informação).

Na construção de modelos atuariais costuma ser boa prática a inserção de variáveis exclusivamente categóricas, com relevância pertinente para a estrutura tarifária a construir. Em modelos multiplicativos - como os que estaremos prestes a construir - estas variáveis categóricas irão estar associadas a fatores tarifários.

A construção destas variáveis categóricas pode ser vista nos anexos relevantes (secção **Modelação – alguns passos prévios**).

Distribuições mais comuns

Na análise exploratória de dados da carteira vimos que a média de sinistros ocorridos por apólice e por ano era igual (ou pelo menos bem próxima) à variância amostral, pelo que, na ausência de sobredispersão parece ser bastante interessante o ajustamento de modelos de Poisson, com parâmetros próximos de zero

(pois a ocorrência de sinistros é um evento raro). Por isso, iremos considerar um modelo de Poisson, para a frequência.

Iremos estimar os parâmetros destas possíveis distribuições, e avaliar a qualidade de ajustamento das mesmas, começando pela distribuição das frequências - a qual se especula que seja uma distribuição de Poisson, com parâmetro dado pelo método da máxima verosimilhança (hipótese nula), versus a negação desta proposição (hipótese alternativa). A estatística de teste será dada, sob a validade de H_0, por:

$$X = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \cap \chi^2_{k-r-1}$$

com $k = 4$ e $r = 1$, sendo O_i a frequência absoluta observada para a i-ésima classe, e E_i a frequência absoluta esperada sob a hipótese nula (distribuição de Poisson). O que se verifica é que a hipótese nula deve ser rejeitada, a qualquer nível de significância minimamente aceitável (pois $p \approx 1.546988 \times 10^{-69} \approx 0$): Porém, esta rejeição não implica que o modelo de regressão de Poisson não seja aplicável ao nosso conjunto de dados.

Modelo de regressão de Poisson

Vimos, em cima, que a hipótese nula é rejeitada, não seguindo o número de sinistros uma distribuição de Poisson. Porém, a verdade é que num modelo de regressão de Poisson não devemos assumir que há uma única distribuição de Poisson, de caráter global, mas sim a existência de várias distribuições de Poisson, de caráter mais individual - uma para cada combinação possível de valores das covariáveis.

Vamos assumir que **var1**, **var2**, **var3**, **var5**, **var6** e **var7** são variáveis que podem dar azo a fatores tarifários de interesse. Nesse caso, estudemos se podemos considerar que, na combinação de níveis que parecem ser mais populares nestes fatores tarifários, estamos perante uma distribuição de Poisson. As hipóteses em teste são similares às descritas anteriormente (num subconjunto de dados mais específico), mas agora $p \approx 0.9377052$, pelo que claramente mantemos a hipótese nula nestas circunstâncias, a qualquer nível de significância razoável:

Tabela 6.3 – Teste à distribuição da frequência de sinistros (cenário específico)

Frequências esperadas (de acordo com a hipótese nula)	Frequências observadas	Frequências esperadas (de acordo com a hipótese nula)	Contribuição aditiva para a estatística de teste
0	374	373.0641	0.0023
1 e mais	6	6.9359	0.1263
Total	380	380	0.2526

Agora, o *p-value* é elevadíssimo, pelo que podemos considerar que (pelo menos) neste conjunto de observações específicas, as frequências seguem uma distribuição de Poisson com um parâmetro λ próprio.

Passemos à construção dos modelos propriamente ditos. Comecemos pela modelação de frequências através de regressões de Poisson, uma vez que se considera que pedidos de indemnização chegam à seguradora através de um processo de Poisson:

```
modelo.freqs.1 <- glm(num_sinistros ~ var1 + var2 +var3 + var4 + var5 + var6 + var7, offset = duracao_risco_anos, family = "poisson", data = frequencias_treino)
summary(modelo.freqs.1)

##
## Call:
## glm(formula = num_sinistros ~ var1 + var2 +var3 +
##     var4 + var5 + var6 + var7,
##     family = "poisson", data = frequencias_treino, offset = duracao_risco_anos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4610  -0.2287  -0.1765  -0.1481   5.0881
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -6.85104    0.30298 -22.612  < 2e-16 ***
## var1categB          0.69374    0.09244   7.505 6.15e-14 ***
## var1categC          -0.50974    0.33745  -1.511 0.130896
## var1categD          0.84325    0.07908  10.663  < 2e-16 ***
## var1categE         0.67853    0.18224   3.723 0.000197 ***
## var1categF    0.65977    0.15094   4.371 1.24e-05 ***
## var1categG          0.65691    0.09905   6.632 3.31e-11 ***
## var1categH          -0.69846    0.32047  -2.179 0.029296 *
## var1categI          -0.68354    0.15466  -4.420 9.89e-06 ***
## var1categJ          0.89390    0.19335   4.623 3.78e-06 ***
## var1categK          0.55011    0.11399   4.826 1.39e-06 ***
## var1categL        0.50893    0.22437   2.268 0.023316 *
## var1categM          0.71211    0.06700  10.628  < 2e-16 ***
## var1categN     -0.01683    0.29368  -0.057 0.954295
## var1categO     -0.13862    0.19291  -0.719 0.472426
## var1categP         0.21870    0.12902   1.695 0.090043 .
## var1categQ          0.03790    0.09255   0.409 0.682187
## var1categR  1.09459    0.11600   9.436  < 2e-16 ***
## var1categS         0.79375    0.15124   5.248 1.54e-07 ***
## var1categT          0.75066    0.12328   6.089 1.14e-09 ***
## var2categB          0.75368    0.22736   3.315 0.000916 ***
## var2categC          1.09381    0.22776   4.802 1.57e-06 ***
## var2categD          1.27657    0.22074   5.783 7.33e-09 ***
## var2categE          1.42537    0.23004   6.196 5.78e-10 ***
## var2categF          1.98685    0.24273   8.185 2.71e-16 ***
## var3categB        -0.12215    0.08033  -1.521 0.128349
## var4categB         0.08182    0.17873   0.458 0.647112
## var4categC         0.08899    0.16907   0.526 0.598648
## var4categD         0.06353    0.16568   0.383 0.701370
## var5categB       0.19109    0.05774   3.309 0.000935 ***
## var5categC       0.03809    0.07809   0.488 0.625695
```

```
## var6categA               -0.32154    0.72403  -0.444 0.656968
## var7categB                0.22910    0.70866   0.323 0.746483
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 18822  on 123448  degrees of freedom
## Residual deviance: 18174  on 123416  degrees of freedom
## AIC: 22786
##
## Number of Fisher Scoring iterations: 7
```

```
drop1(modelo.freqs.1)
```

```
## Single term deletions
##
## Model:
## num_sinistros ~ var1 + var2 +var3 +
##     var4 + var5 + var6 + var7
##                 Df Deviance   AIC
## <none>                18175 22786
## var1          19    18562 23135
## var2       5    18302 22903
## var3    1    18177 22786
## var4    3    18175 22780
## var5   2    18186 22793
## var6            1    18175 22784
## var7            1    18175 22784
```

Aqui chegados, importa abordar em maior profundidade de que forma o R parametriza variáveis categóricas. Por predefinição, o R considera como categoria base (nível base do factor associado) a primeira que consegue identificar, por ordem alfabética; porém, esta é apenas uma predefinição, a qual não impede escolhas próprias e deliberadas.

Tendo em conta o que foi feito neste anexo, podemos indicar que o R insere uma variável categórica no modelo linear generalizado codificando automaticamente cada um dos seus possíveis valores através de variáveis binárias, sem necessidade de intervenção por parte do utilizador. A única exceção é mesmo a categoria base, a qual é "absorvida" no *intercept*, não aparecendo de forma explícita no *output* do R.

Iremos, neste processo de construção de modelos, implementar uma abordagem através da qual começamos com todas as variáveis que nos façam sentido dentro do modelo, removendo-as uma a uma, se justificado através do AIC (quanto menor, melhor), e se **não** existirem bons motivos lógicos ou de negócio que impeçam a sua remoção.

Poderá ser interessante a remoção de **var4**.

```
modelo.freqs.2 <- glm(num_sinistros ~ var1 + var2 + var3 + var5 + var6 + var7, offset = duracao_risco_a
nos, family = "poisson", data = frequencias_treino)
summary(modelo.freqs.2)
```

```
##
## Call:
## glm(formula = num_sinistros ~ var1 + var2 + var3 +
##     var5 + var6 + var7, family = "poisson",
##     data = frequencias_treino, offset = duracao_risco_anos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4614  -0.2290  -0.1759  -0.1488   5.0860
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -6.75620    0.22326 -30.262  < 2e-16 ***
## var1categB             0.69147    0.09207   7.511 5.88e-14 ***
## var1categC            -0.51057    0.33741  -1.513 0.130224
## var1categD             0.84089    0.07839  10.727  < 2e-16 ***
## var1categE          0.67477    0.18167   3.714 0.000204 ***
## var1categF     0.65699    0.15055   4.364 1.28e-05 ***
## var1categG          0.65595    0.09890   6.633 3.29e-11 ***
## var1categH           -0.69819    0.32044  -2.179 0.029340 *
## var1categI            -0.68405    0.15463  -4.424 9.70e-06 ***
## var1categJ             0.89033    0.19307   4.611 4.00e-06 ***
## var1categK             0.54721    0.11367   4.814 1.48e-06 ***
## var1categL        0.50710    0.22426   2.261 0.023742 *
## var1categM             0.71107    0.06685  10.638  < 2e-16 ***
## var1categN      -0.01719    0.29355  -0.059 0.953308
## var1categO     -0.13895    0.19291  -0.720 0.471366
## var1categP          0.21670    0.12871   1.684 0.092246 .
## var1categQ          0.03740    0.09253   0.404 0.686042
## var1categR  1.09297    0.11568   9.448  < 2e-16 ***
## var1categS         0.79114    0.15083   5.245 1.56e-07 ***
## var1categT             0.74786    0.12282   6.089 1.13e-09 ***
## var2categB          0.75160    0.22613   3.324 0.000888 ***
## var2categC          1.08471    0.22406   4.841 1.29e-06 ***
## var2categD          1.25177    0.21471   5.830 5.54e-09 ***
## var2categE          1.40618    0.22590   6.225 4.82e-10 ***
## var2categF          1.95918    0.23603   8.301  < 2e-16 ***
## var3categB -0.12571    0.07853  -1.601 0.109434
## var5categB       0.19076    0.05707   3.342 0.000831 ***
## var5categC       0.02797    0.07211   0.388 0.698140
## var6categA              -0.38414    0.71232  -0.539 0.589694
## var7categB               0.23539    0.70854   0.332 0.739728
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 18822  on 123448  degrees of freedom
## Residual deviance: 18175  on 123419  degrees of freedom
## AIC: 22780
```

```
##
## Number of Fisher Scoring iterations: 7
```

```
anova(modelo.freqs.1, modelo.freqs.2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: num_sinistros ~ var1 + var2 +var3 +
##     var4 + var5 + var6 + var7
## Model 2: num_sinistros ~ var1 + var2 +var3 +
##     var5 + var6 + var7
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1    123416      18175
## 2    123419      18175 -3 -0.32946   0.9544
```

```
drop1(modelo.freqs.2)
```

```
## Single term deletions
##
## Model:
## num_sinistros ~ var1 + var2 +var3 +
##     var5 + var6 + var7
##                 Df Deviance   AIC
## <none>                18175 22780
## var1          19    18570 23137
## var2       5    18353 22948
## var3    1    18178 22781
## var5   2    18186 22788
## var6           1    18175 22778
## var7           1    18175 22778
```

De seguida poderíamos optar por remover **var6** e, de seguida, **var7** (em qualquer ordem). Porém, quando removemos uma, a outra passa a ser estatisticamente significativa - talvez pelo facto de estarem muito correlacionadas ($\rho \approx 0.9987$). Porém, também devemos salientar que estas variáveis são *irmãs*, pois resultam originalmente de uma outra, entretanto eliminada, pelo que não parece fazer muito sentido remover uma e manter outra, tendo também em conta que esta escolha é algo arbitrária, e age num modelo que já nos parece, dadas as condições, suficientemente bom.

Por último, e mais importante, não faria logicamente sentido remover **var6** do modelo, uma vez que isso resultaria num modelo com um coeficiente negativo para **var7**, valor que economicamente não faz nenhum sentido (e que, adicionalmente, dá origem a uma estrutura tarifária incompleta).

```
summary(glm(num_sinistros ~ var1 + var2 +var3 + var5 + var7, offset = duracao_risco_anos, family = "p
oisson", data = frequencias_treino))
```

```
##
## Call:
## glm(formula = num_sinistros ~ var1 + var2 +var3 +
##     var5 + var7, family = "poisson", data = frequencias_treino,
##     offset = duracao_risco_anos)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4615  -0.2291  -0.1761  -0.1489   5.0853
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -6.76008    0.22316 -30.292  < 2e-16 ***
## var1categB           0.69150    0.09206   7.511 5.86e-14 ***
## var1categC          -0.51067    0.33741  -1.514 0.130148
## var1categD           0.84040    0.07839  10.721  < 2e-16 ***
## var1categE          0.67419    0.18167   3.711 0.000206 ***
## var1categF    0.65647    0.15054   4.361 1.30e-05 ***
## var1categG          0.65584    0.09889   6.632 3.32e-11 ***
## var1categH          -0.69781    0.32043  -2.178 0.029428 *
## var1categI          -0.68430    0.15463  -4.426 9.62e-06 ***
## var1categJ           0.88947    0.19306   4.607 4.08e-06 ***
## var1categK           0.54674    0.11366   4.810 1.51e-06 ***
## var1categL        0.50671    0.22425   2.260 0.023851 *
## var1categM           0.71070    0.06684  10.633  < 2e-16 ***
## var1categN      -0.01779    0.29355  -0.061 0.951670
## var1categO     -0.13916    0.19291  -0.721 0.470699
## var1categP          0.21678    0.12870   1.684 0.092126 .
## var1categQ          0.03765    0.09253   0.407 0.684093
## var1categR  1.09237    0.11567   9.444  < 2e-16 ***
## var1categS         0.79070    0.15083   5.242 1.59e-07 ***
## var1categT           0.74740    0.12281   6.086 1.16e-09 ***
## var2categB          0.75513    0.22602   3.341 0.000835 ***
## var2categC          1.08897    0.22393   4.863 1.16e-06 ***
## var2categD          1.25477    0.21465   5.846 5.04e-09 ***
## var2categE          1.41049    0.22577   6.248 4.17e-10 ***
## var2categF          1.96346    0.23591   8.323  < 2e-16 ***
## var3categB -0.12402    0.07846  -1.581 0.113942
## var5categB       0.19147    0.05705   3.356 0.000791 ***
## var5categC       0.02577    0.07202   0.358 0.720427
## var7categB              -0.14493    0.08181  -1.772 0.076461 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 18822  on 123448  degrees of freedom
## Residual deviance: 18175  on 123420  degrees of freedom
## AIC: 22778
##
## Number of Fisher Scoring iterations: 7
```

Uma outra escolha, talvez mais adequada em caso de sobredispersão, mas definitivamente mais complexa e originalmente não inserida na teoria dos modelos lineares generalizados, passaria pelo recurso a uma distribuição Binomial Negativa.

Modelação da severidade de sinistros

Distribuições mais comuns

Na análise exploratória de dados, vimos:

- Que a distribuição dos custos base por sinistro carateriza-se pela sua elevada dispersão e pela sua elevada assimetria;
- Que o logaritmo neperiano dos custos base com sinistros parece seguir aproximadamente uma distribuição Normal.

Por isso, iremos considerar modelos Gama e log-Normal para a severidade, fazendo com que estes compitam entre si.

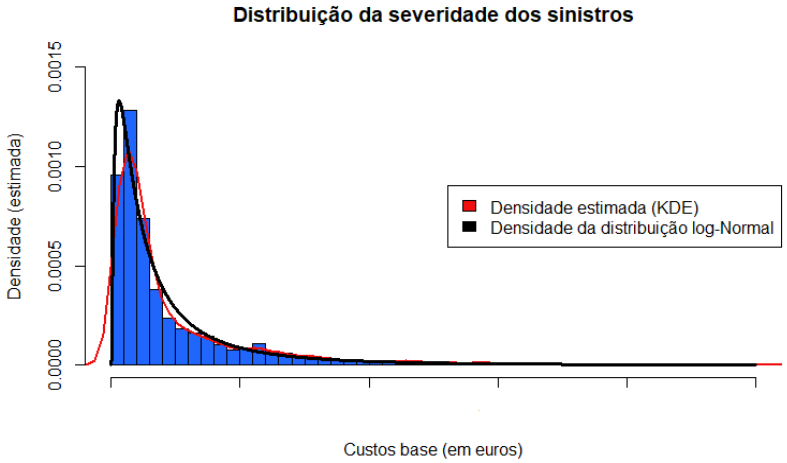

Figura 6.1– Distribuição real/empírica da severidade dos sinistros vs distribuição lognormal

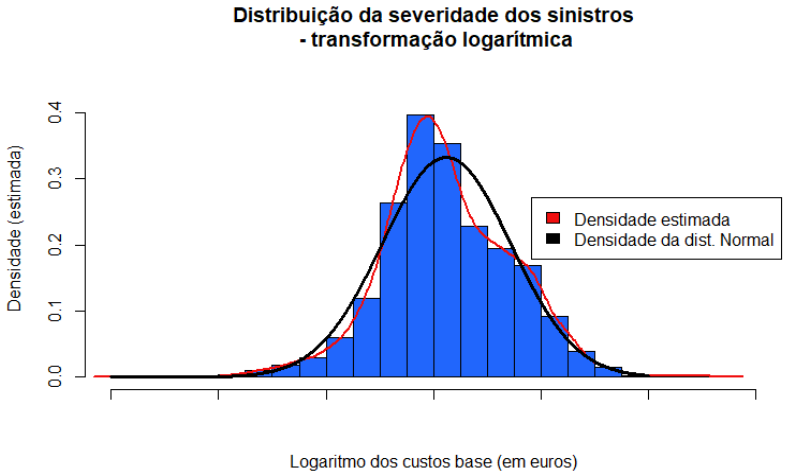

Figura 6.2– Distribuição real/empírica da severidade dos sinistros, após transformação logarítmica

Para a distribuição Gama, e considerando o método dos momentos (mais acessível do que o método de máxima verosimilhança, o qual exige uma resolução numérica) foi possível obter o seguinte gráfico:

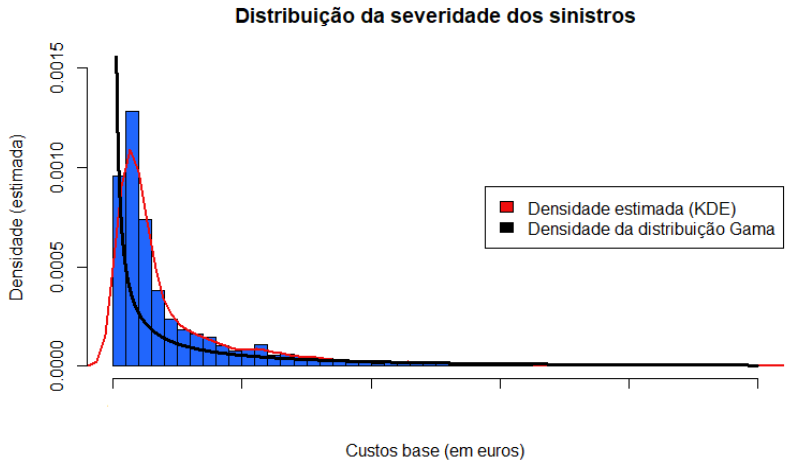

Figura 6.3 - Distribuição empírica da severidade dos sinistros vs distribuição Gama

Verificamos que a aplicação da distribuição log-Normal parece ser bem mais promissora, mesmo não havendo igualdade de condições - dado que os parâmetros da distribuição Gama não foram estimados através do método de máxima verosimilhança.

Comecemos então a pensar em ajustar modelos de regressão aos dados.

Modelo #1: regressão linear múltipla, com dados logaritmizados

Comecemos pelo recurso à distribuição log-Normal, o que implica o recurso ao modelo clássico de regressão linear, mais simples e com melhores propriedades estatísticas (isto é, mais robustas), depois de uma transformação logarítmica prévia à variável resposta.

Seguindo a mesma abordagem de remoção de variáveis uma a uma,

```
modelo.custos.lognormal.1 <- lm(log(custo_base) ~ var1 + var2 +var3 + var4 + var5 + var6 + var7, data = custos_treino)
summary(modelo.custos.lognormal.1)

##
## Call:
## lm(formula = log(custo_base) ~ var1 + var2 +var3 +
##     var4 + var5 + var6 + var7,
##     data = custos_treino)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2001 -0.6636  0.0201  0.6713  4.6856
##
## Coefficients:
```

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.349620   0.381534  16.642   <2e-16 ***
## var1categB         -0.184534   0.107107  -1.723   0.0850 .
## var1categC          0.031230   0.369156   0.085   0.9326
## var1categD          0.001883   0.091546   0.021   0.9836
## var1categE         -0.203140   0.238362  -0.852   0.3942
## var1categF    0.005974   0.158234   0.038   0.9699
## var1categG         -0.115397   0.119249  -0.968   0.3333
## var1categH          0.154364   0.324012   0.476   0.6338
## var1categI          0.113918   0.159925   0.712   0.4763
## var1categJ          0.048254   0.231055   0.209   0.8346
## var1categK         -0.047490   0.134638  -0.353   0.7243
## var1categL       -0.198465   0.312604  -0.635   0.5256
## var1categM          -0.072758   0.077009  -0.945   0.3449
## var1categN       0.226009   0.303016   0.746   0.4558
## var1categO      0.216616   0.234381   0.924   0.3555
## var1categP         -0.022665   0.149128  -0.152   0.8792
## var1categQ          0.175299   0.106815   1.641   0.1009
## var1categR  0.047518   0.131329   0.362   0.7175
## var1categS        -0.010623   0.186824  -0.057   0.9547
## var1categT          -0.132825   0.140795  -0.943   0.3456
## var2categB          0.098088   0.244197   0.402   0.6880
## var2categC          0.069389   0.249210   0.278   0.7807
## var2categD          0.158213   0.244882   0.646   0.5183
## var2categE          0.248638   0.254107   0.978   0.3279
## var2categF          0.216491   0.272229   0.795   0.4265
## var3categB  0.207871   0.095240   2.183   0.0292 *
## var4categB        0.019813   0.225453   0.088   0.9300
## var4categC       -0.070749   0.208367  -0.340   0.7342
## var4categD        0.011296   0.202245   0.056   0.9555
## var5categB       0.099578   0.068130   1.462   0.1440
## var5categC       0.128241   0.087599   1.464   0.1433
## var6categA              -0.606578   0.836762  -0.725   0.4686
## var7categB              -0.183002   0.816290  -0.224   0.8226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.145 on 2357 degrees of freedom
## Multiple R-squared:  0.1036, Adjusted R-squared:  0.09146
## F-statistic: 8.515 on 32 and 2357 DF,  p-value: < 2.2e-16
```

drop1(modelo.custos.lognormal.1)

```
## Single term deletions
##
## Model:
## log(custo_base) ~ var1 + var2 +var3 +
##     var4 + var5 + var6 + var7
##              Df Sum of Sq    RSS    AIC
## <none>                    3092.5 681.84
## var1         19   20.2786 3112.7 659.46
```

```
## var2      5    6.2240 3098.7 676.65
## var3   1    6.2502 3098.7 684.67
## var4   3    2.1013 3094.6 677.47
## var5  2    4.1553 3096.6 681.05
## var6          1    0.6895 3093.2 680.37
## var7          1    0.0659 3092.5 679.89
```

Para verificar se as diferenças entre o modelo inicial (com p parâmetros) e o sub-modelo a partir deste construído (com $p - q$ parâmetros) são estatisticamente significativas, e sendo S o conjunto de parâmetros presentes no modelo global, mas não no sub-modelo, podemos testar a seguinte **hipótese linear**:

$$H_0: \beta_j = 0, \forall j \in S \; vs \; H_1: \exists j \in S: \beta_j \neq 0$$

A estatística de teste será:

$$F = \frac{\frac{SSE_{reduzido} - SSE_{global}}{q}}{\frac{SSE_{global}}{n - p}}$$

Sob a validade de H_0, esta estatística tem distribuição F com q graus de liberdade no numerador e $n - p$ no denominador.

Foi seguido de forma manual um ciclo de ponderação, remoção de uma variável e posterior ajustamento do modelo, até se considerar (devido a motivos lógicos, estatísticos e/ou de negócio) que mais nenhuma variável deve ser removida. Assim sendo, procedeu-se à remoção sequencial de:

1. **var1**;
2. **var4**;
3. **var2**;

para, mais uma vez, deixar **var6** e **var7** juntas - sendo mantidas ou removidas em conjunto. Chegou-se então ao seguinte modelo:

```
modelo.custos.lognormal.2 <- lm(log(custo_base) ~ var3 + var5 + var6 + var7, data = custos_treino)
summary(modelo.custos.lognormal.2)

##
## Call:
## lm(formula = log(custo_base) ~ var3 + var5 +
##     var6 + var7, data = custos_treino)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1350 -0.6619  0.0009  0.6762  4.7329
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.46433    0.04065 159.039   <2e-16 ***
## var3categB  0.23050    0.09047   2.548   0.0109 *
## var5categB       0.06953    0.06413   1.084   0.2784
```

```
## var5categC       0.14974    0.07492   1.999   0.0458 *
## var6categA              -0.66000    0.81484  -0.810   0.4180
## var7categB              -0.15299    0.81172  -0.188   0.8505
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.144 on 2384 degrees of freedom
## Multiple R-squared:  0.09488,    Adjusted R-squared:  0.09298
## F-statistic: 49.98 on 5 and 2384 DF,  p-value: < 2.2e-16
```

```
anova(modelo.custos.lognormal.1, modelo.custos.lognormal.2)
```

```
## Analysis of Variance Table
##
## Model 1: log(custo_base) ~ var1 + var2 +var3 +
##     var4 + var5 + var6 + var7
## Model 2: log(custo_base) ~ var3 + var5 + var6 +
##     var7
##   Res.Df    RSS  Df Sum of Sq     F Pr(>F)
## 1   2357 3092.5
## 2   2384 3122.7 -27   -30.184 0.852 0.6842
```

```
drop1(modelo.custos.lognormal.2)
```

```
## Single term deletions
##
## Model:
## log(custo_base) ~ var3 + var5 + var6 +
##     var7
##                Df Sum of Sq    RSS    AIC
## <none>                      3122.7 651.06
## var3   1    8.5025 3131.2 655.55
## var5  2    5.5659 3128.2 651.31
## var6           1    0.8593 3123.5 649.71
## var7           1    0.0465 3122.7 649.09
```

Modelo #2: regressão Gama

Repetindo os raciocínios anteriores para um modelo de regressão Gama, com função de ligação logarítmica, teremos inicialmente:

```
modelo.custos.gama.1 <- glm(custo_base ~ var1 + var2 +var3 + var4 + var5 + var6 + var7, family = Gamma(link = "log"), data = custos_treino)
summary(modelo.custos.gama.1)
```

```
##
## Call:
## glm(formula = custo_base ~ var1 + var2 +var3 +
##     var4 + var5 + var6 + var7,
```

```
##     family = Gamma(link = "log"), data = custos_treino)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5704  -1.0935  -0.5322   0.0718   9.4898
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.514751   0.666071  11.282   <2e-16 ***
## var1categB          -0.225465   0.186985  -1.206   0.2280
## var1categC          -0.257012   0.644462  -0.399   0.6901
## var1categD           0.141107   0.159818   0.883   0.3774
## var1categE          -0.533392   0.416126  -1.282   0.2000
## var1categF          -0.377120   0.276240  -1.365   0.1723
## var1categG          -0.369589   0.208181  -1.775   0.0760 .
## var1categH          -0.397234   0.565651  -0.702   0.4826
## var1categI          -0.005672   0.279192  -0.020   0.9838
## var1categJ          -0.362294   0.403368  -0.898   0.3692
## var1categK          -0.165204   0.235047  -0.703   0.4822
## var1categL          -0.447151   0.545734  -0.819   0.4127
## var1categM          -0.282341   0.134441  -2.100   0.0358 *
## var1categN           0.597546   0.528996   1.130   0.2588
## var1categO          -0.014899   0.409175  -0.036   0.9710
## var1categP          -0.139375   0.260343  -0.535   0.5925
## var1categQ           0.023243   0.186474   0.125   0.9008
## var1categR -0.307323   0.229271  -1.340   0.1802
## var1categS          -0.283404   0.326153  -0.869   0.3850
## var1categT          -0.410337   0.245796  -1.669   0.0952 .
## var2categB          -0.224046   0.426312  -0.526   0.5993
## var2categC          -0.229085   0.435063  -0.527   0.5986
## var2categD          -0.080428   0.427507  -0.188   0.8508
## var2categE           0.121423   0.443612   0.274   0.7843
## var2categF           0.005697   0.475249   0.012   0.9904
## var3categB  0.063197   0.166268   0.380   0.7039
## var4categB           0.135278   0.393590   0.344   0.7311
## var4categC          -0.074783   0.363761  -0.206   0.8371
## var4categD          -0.034597   0.353074  -0.098   0.9219
## var5categB          -0.012681   0.118940  -0.107   0.9151
## var5categC           0.019207   0.152928   0.126   0.9001
## var6categA          -1.237535   1.460795  -0.847   0.3970
## var7categB           0.266064   1.425056   0.187   0.8519
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 3.998708)
##
##     Null deviance: 3738.2  on 2389  degrees of freedom
## Residual deviance: 3124.4  on 2357  degrees of freedom
## AIC: 37829
##
## Number of Fisher Scoring iterations: 9
```

```
drop1(modelo.custos.gama.1)

## Single term deletions
##
## Model:
## custo_base ~ var1 + var2 +var3 +
##     var4 + var5 + var6 + var7
##                 Df Deviance   AIC
## <none>               3124.4 37829
## var1          19   3211.1 37813
## var2       5   3147.7 37825
## var3    1   3125.0 37827
## var4    3   3131.5 37825
## var5  2   3124.6 37825
## var6           1   3128.4 37828
## var7           1   3124.6 37827
```

Procedeu-se, de forma justificada, à remoção sequencial de

1. **var1**;
2. **var5**;
3. **var4**;
4. **var2**;
5. **var3**;

Obteve-se então o seguinte modelo:

```
modelo.custos.gama.2 <- glm(custo_base ~ var6 + var7, family = Gamma(link = "log"), data = custos_tre
ino)
summary(modelo.custos.gama.2)

##
## Call:
## glm(formula = custo_base ~ var6 + var7, family = Gamma(link = "log"),
##     data = custos_treino)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4564  -1.1116  -0.5532   0.0172  11.1405
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.30176    0.06258 116.684   <2e-16 ***
## var6categA   -1.44214    1.69139  -0.853    0.394
## var7categB    0.46573    1.69074   0.275    0.783
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 5.709354)
##
##     Null deviance: 3738.2  on 2389  degrees of freedom
```

```
## Residual deviance: 3248.7  on 2387  degrees of freedom
## AIC: 37881
##
## Number of Fisher Scoring iterations: 7
```

```
anova(modelo.custos.gama.1, modelo.custos.gama.2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: custo_base ~ var1 + var2 + var3 +
##     var4 + var5 + var6 + var7
## Model 2: custo_base ~ var6 + var7
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1      2357     3124.4
## 2      2387     3248.7 -30  -124.23   0.4121
```

```
drop1(modelo.custos.gama.2)
```

```
## Single term deletions
##
## Model:
## custo_base ~ var6 + var7
##          Df Deviance   AIC
## <none>       3248.7 37881
## var6  1   3255.8 37880
## var7  1   3249.2 37879
```

De notar, pelos motivos já considerados, que **var6** e **var7** permanecem juntas no modelo; apesar de podermos remover apenas uma, a remoção simultânea de ambas faz com que o modelo obtido - agora nulo - seja significativamente diferente do inicial, estatisticamente falando.

A função de ligação nesta regressão é a logarítmica, porque pretendemos obter um modelo de caráter multiplicativo para a estrutura tarifária e porque a função de ligação canónica (inversa, neste caso) **não evita a obtenção de valores negativos**, apesar de cada custo com sinistros ter de ser positivo.

Seleção, diagnóstico e validação de modelos (nos conjuntos de treino e de validação)

Tendo sido construídos modelos de regressão de Poisson, de regressão linear com transformação logarítmica prévia e de regressão Gama, passa agora a ser necessário aferir o grau de adequação e a qualidade de ajustamento dos mesmos. Por ser o mais simples e beneficiar de propriedades mais agradáveis, o modelo de regressão linear será aquele que mais se prestará a diagnósticos. Por isso, executemos:

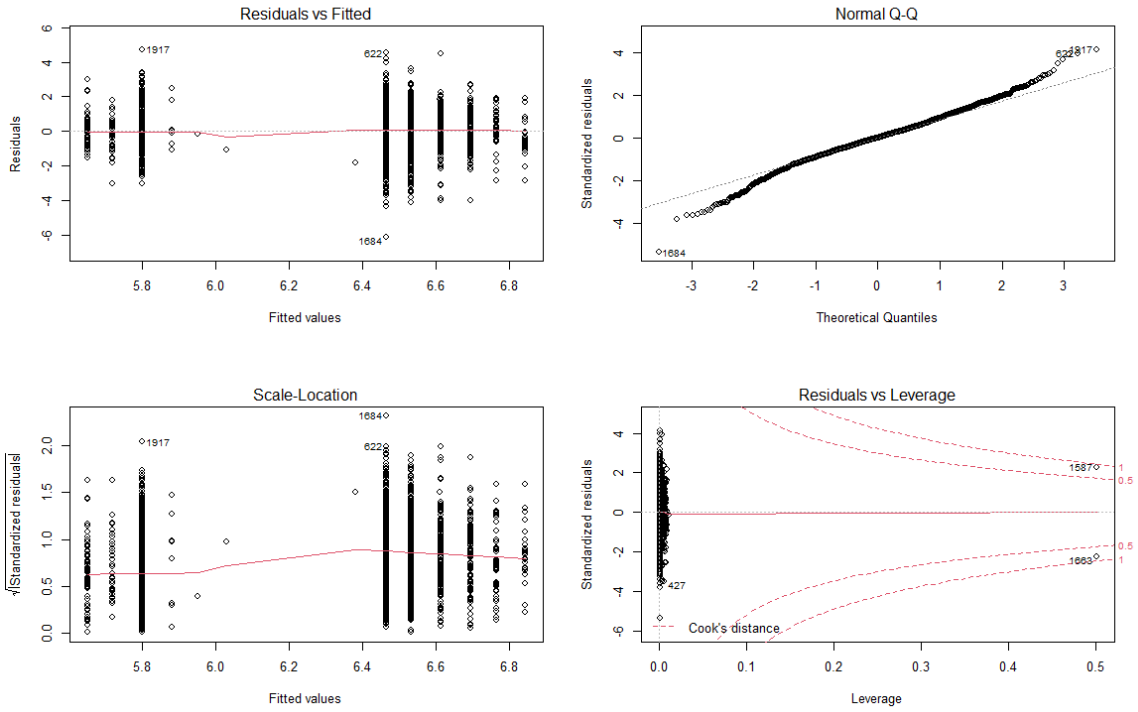

Figura 6.4 - Diagnóstico do modelo de regressão linear com transformação logarítmica para a variável resposta

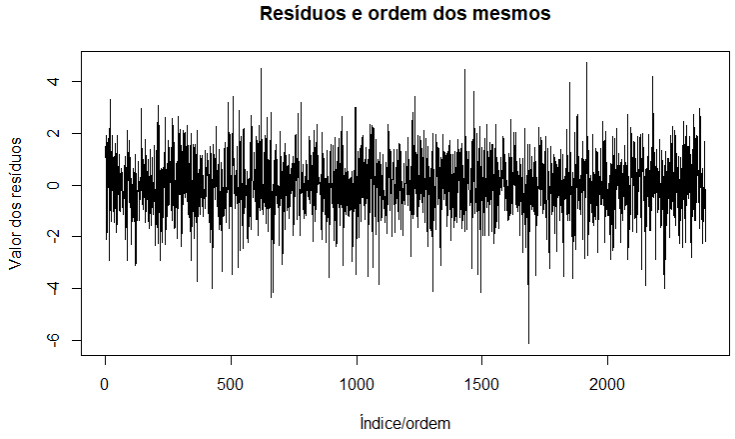

Figura 6.5 - Gráfico dos resíduos da regressão linear vs ordem dos mesmos

Podemos observar, em cada gráfico, que:

1. Os resíduos e os valores ajustados (*fitted values*) não parecem exibir nenhum padrão, o que valida o pressuposto de linearidade dos parâmetros;
2. O QQ-plot indica-nos que a distribuição dos resíduos parece ser *suficientemente normal* - exceto nas caudas, se bem que não esperávamos um ajuste perfeito, sendo este tema analisado mais em detalhe de seguida;

3. Os resíduos parecem estar dispostos numa banda horizontal, e a a linha a vermelho parece ser *suficientemente reta*, o que valida o pressuposto de homoscedasticidade;
4. Parecem haver poucas observações influentes, sendo que a remoção das observações 1587 e 1663 em **custos_treino** não parece causar grandes alterações nos modelos construídos;
5. Na relação entre os resíduos e a ordem destes, não parecem haver padrões, pelo que não parece haver autocorrelação entre os mesmos, o que valida o pressuposto de que os erros da regressão são ruído branco.

Nota: Estatisticamente falando, *leverage* ou, em português, (efeito de) alavanca, ocorre quando estamos perante observações onde pelo menos uma covariável apresenta valores extremos, em comparação com a sua média. Estas observações podem, em certos casos, exercer um efeito considerável nas estimativas de coeficientes de regressão.

De notar que a consideração de variáveis de natureza categórica faz com que existam apenas 13 valores ajustados distintos.

Relativamente aos resíduos, estes têm média nula (o que acontece sempre que usamos o método dos mínimos quadrados para estimar regressões com termo constante), e mediana também próxima de zero (sendo aproximadamente 0.000921). O seu desvio padrão é de cerca de 1.14, estando a assimetria próxima de zero (sendo aproximadamente -0.194). Tudo isto nos aproxima à hipótese de Normalidade dos erros, exceto talvez a curtose, cujo coeficiente é de 4.29 sendo que, sob Normalidade, estaríamos à espera de que este valor fosse igual a 3.

A nível visual, o pressuposto de Normalidade dos erros (via resíduos) parece ganhar cada vez mais força:

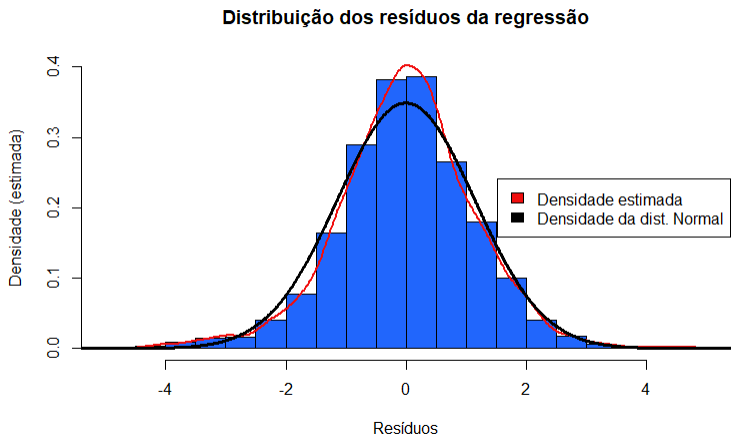

Figura 6.6 - Distribuição dos resíduos associados ao modelo de regressão linear

Para o caso de ser considerada necessária uma análise mais rigorosa, podemos sempre efetuar um teste do qui-quadrado às seguintes hipóteses:

H_0: Os resíduos seguem uma distribuição Normal com parâmetros dados pelos estimadores de máxima verosimilhança *vs* H_1: Os resíduos **não** seguem uma distribuição Normal, pelo menos não com tais parâmetros

com a estatística de teste já vista,. Somos levados a rejeitar H_0 (pois $p \approx 4.227062 \times 10^{-6}$), mas esta rejeição pode ser motivada pelas elevadas contagens, pelo que iremos manter a hipótese de Normalidade.

Relativamente aos *hat values*, temos um valor máximo de 0.5015622, acima do limite de 0.2 considerado como regra prática, mas como os resíduos aparentam ter uma distribuição Normal (e, portanto, simétrica), não iremos considerar que hajam problemas aqui (até porque menos de 0.1% destes *hat values* são superiores a 0.2).

Relativamente aos restantes modelos (lineares generalizados), a realização de diagnósticos é mais complicada. Quanto muito, poderemos executar **glm.diag.plots**(modelo.custos.gama.2) e **glm.diag.plots**(modelo.freqs.2), obtendo-se:

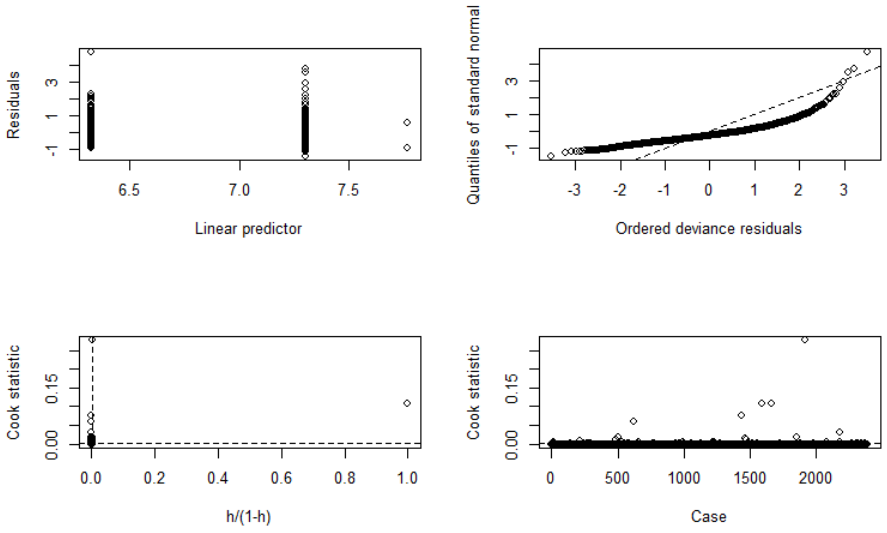

Figura 6.7 - Diagnóstico do modelo de regressão Gama para modelação de severidades

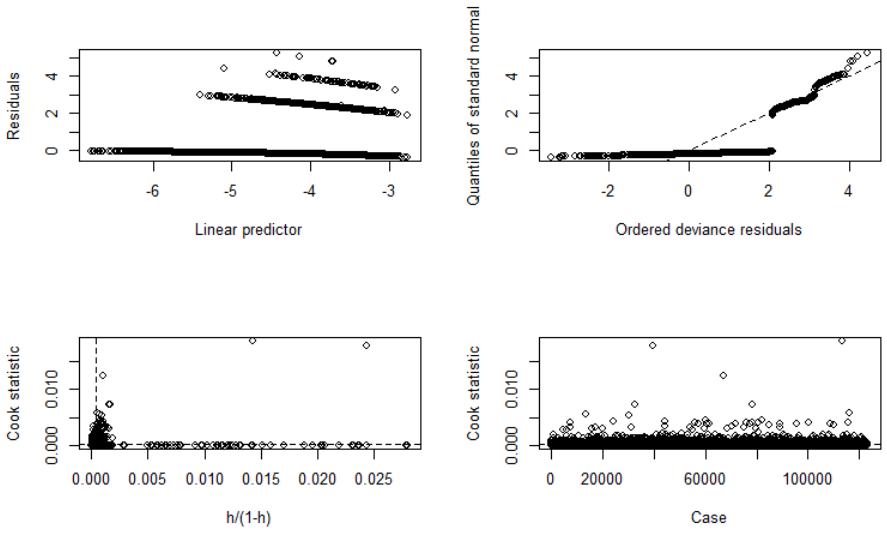

Figura 6.8 - Diagnóstico do modelo de regressão de Poisson para modelação de frequências

De notar que, na regressão Gama, a consideração de apenas duas variáveis de natureza categórica faz com que existam apenas 3 valores ajustados distintos. Isto acontece pois, dos quatro pares de valores possíveis, um é na verdade inadmissível e outro extremamente raro.

Já ao nível do modelo de regressão de Poisson, a análise do gráfico no canto superior esquerdo revela a existência de "linhas" de observações algo paralelas entre si, que tendem a decair à medida que o preditor linear (e, como consequência, a sua transformação exponencial) aumenta. Para além disso, as frequências previstas tendem a ser diminutas e, por outro lado, apólices nas quais se verificam sinistros são minoritárias. Por isso, sem surpresas, a maioria dos resíduos é negativa (pois estamos a falar de apólices com 0 sinistros observados e com um número positivo de sinistros esperados ou previstos), e a cada "escalão" neste gráfico corresponde a um grupo de apólices com um dado número de sinistros por ano comum entre si.

Em qualquer destes modelos lineares generalizados, notamos ainda que:

1. Apesar de os *deviance residuals* não parecerem ser Normais, isto não faz com que o modelo seja inválido;
2. Nos últimos dois gráficos (em cada caso), não parecem haver observações excessivamente influentes.

Ao nosso dispor temos ainda outras abordagens que nos permitem validar estes modelos. Por exemplo, vejamos, para cada variável preditiva inserida nos modelos, se os valores das variáveis resposta previstos pelos mesmos estão efetivamente próximos dos observados.

O modelo de Poisson considerado para as frequências parece bastante adequado, uma vez que capta as tendências observadas nos números de sinistros efetivamente observados (linha a negro), tende a estimar (linha a vermelho) bem estes valores observados, e comete erros de sobrestimação (os quais não deixam de ser erros, que são ainda assim preferíveis a subestimar a ocorrência de azares).

Nota: Dos gráficos de seguida mostrados, apenas se exemplifica a obtenção do primeiro, dado todos os restantes serem similares. Adicionalmente, nos seguintes gráficos:

- A linha a vermelho representa as frequências previstas pelo modelo de Poisson;
- A linha negra representa as frequências efetivamente verificadas.

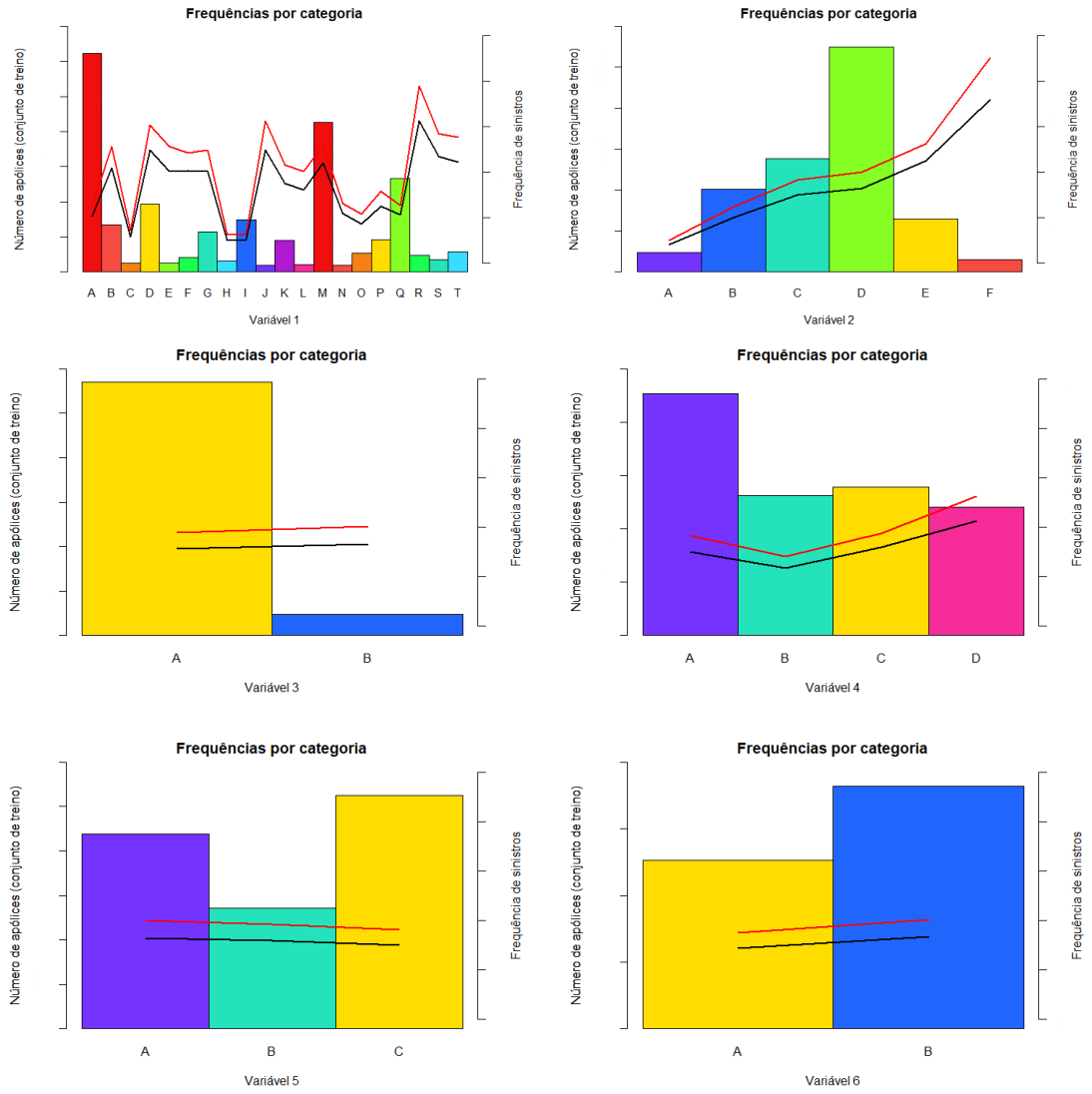

Figura 6.1 - Valores observados vs previstos pelo modelo de regressão de Poisson para a frequência de sinistros por apólice, para cada valor de diversas variáveis categóricas

Já ao nível dos custos, os modelos finais obtidos apresentam mais dificuldades, sobretudo ao nível das variáveis 1 e 4, seja qual for a abordagem escolhida.

Nota: Nos gráficos de seguida mostrados:

- A linha a vermelho representa as severidades previstas pelo modelo lognormal;
- A linha a azul representa as severidades previstas pelo modelo Gama;
- A linha negra representa as severidades efetivamente verificadas.

No último gráfico, só é visível uma linha, pois as linhas azul e negra coincidem.

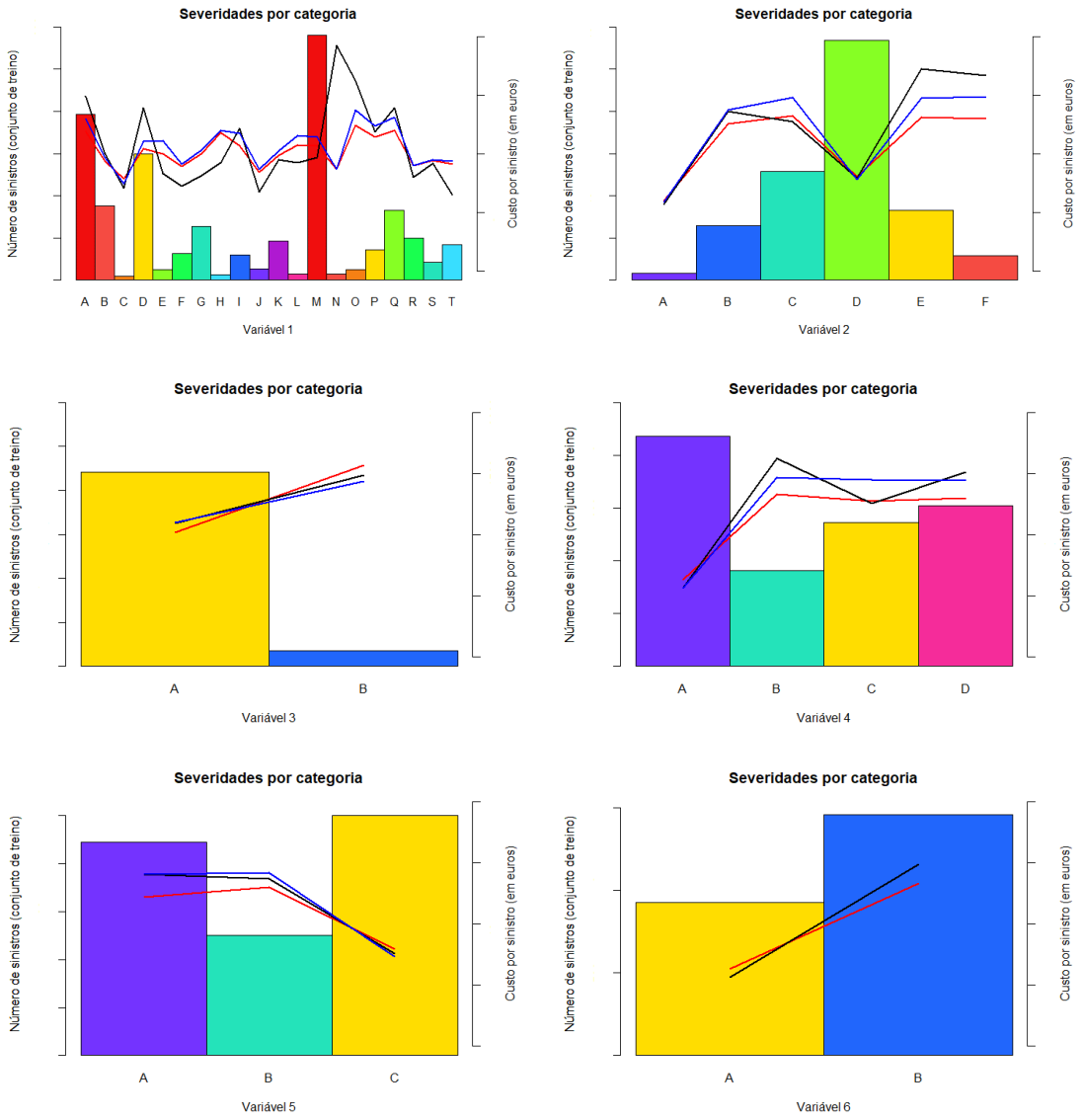

Figura 6.2 - Valores observados vs previstos pelo modelo de regressão Gama (linha azul) e pelo modelo de regressão linear com transformação logarítmica (linha vermelha) para a severidade de sinistros por apólice, para cada valor de diversas variáveis categóricas

Temos, ao nível da qualidade de ajustamento, um R^2 (no modelo log-Normal, de regressão linear múltipla) de 0.0949, e um R^2_{adj} de 0.0930. Isto significa que este modelo não explica mais de 10% da variação presente nos dados.

Já nos (restantes) modelos lineares generalizados, os *pseudo-*R^2 de McFadden indicam-nos que os modelos de regressão Gama e Poisson explicam cerca de 13.09% e 3.44% da *deviance* presente nos dados nos quais foram ajustados, respetivamente. De acordo com este indicador, o modelo Gama sai relativamente valorizado, ao contrário do modelo de Poisson.

Já ao nível de medidas de erro, podemos recorrer à função **mod.error.measures**, por nós criada e presente em **mod.R**, não sem antes importar os conjuntos de validação e de teste, tanto em **frequencias** como em **custos**, e efetuar as operações e transformações já realizadas no conjunto de treino (as quais podem

ser consultadas nos anexos deste trabalho). Mas antes, procuremos definir estas medidas de erro. Seja o erro de estimação associado a uma dada observação dado por

$$e_i = y_i - \hat{y}_i$$

O *mean squared error* corresponde, em Português, ao erro quadrático médio, sendo portanto dado por:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

O *root mean squared error* corresponde à raíz quadrada do erro quadrático médio:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$$

Já o *mean absolute error* corresponde ao erro absoluto médio e é dado por:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |e_i|$$

No modelo de frequências (regressão de Poisson), verificamos mais uma vez que o modelo de frequências é *bem-comportado*, uma vez que generaliza relativamente bem as suas previsões:

Tabela 6.4 - Medidas de erro associadas ao modelo de regressão de Poisson, nos conjuntos de treino, validação e teste

Conjunto	MSE	RMSE	MAE
Treino	0.02035723	0.1426788	0.03713415
Validação	0.02165542	0.1471578	0.0380897
Teste	0.02170138	0.1473139	0.0377205

Já ao nível dos custos, e começando pelo modelo de regressão Gama:

Tabela 6.5 - Medidas de erro associadas ao modelo de regressão Gama, nos conjuntos de treino e validação

Conjunto	MSE	RMSE	MAE
Treino	7350396	2711.161	989.4802
Validação	7758125	2785.341	849.7824

Ao nível do modelo log-Normal (modelo clássico de regressão linear, com transformação logarítmica prévia), temos:

Tabela 6.1 - Medidas de erro associadas ao modelo de regressão linear com transformação logarítmica, nos conjuntos de treino e validação

Conjunto	MSE	RMSE	MAE
Treino	5156971	2270.896	966.07
Validação	5563692	2358.748	833.1409

Com base nestes resultados, verificamos que o modelo log-Normal está associado a menores erros do que o modelo de regressão Gama.

Apesar do modelo Gama ser mais simples (pois possui menos covariáveis), esta parcimónia não é muito relevante, uma vez que o modelo de custo log-Normal possui variáveis que podem também estar inseridas no modelo de frequências (de natureza mais complexa). Adicionalmente, a avaliação e diagnóstico de modelos é bem mais acessível no caso da distribuição log-Normal, uma vez que mediante transformação logarítmica dos dados podemos enquadrar a nossa análise no modelo clássico de regressão linear por nós preferido, pelas razões anteriormente vistas. Por último, a distribuição log-Normal tende, na prática, a ser a mais adequada na modelação de custos com sinistros em seguros multirrisco habitação, como acabamos de ver nas medidas de erro. Por estes motivos, iremos escolher o modelo log-Normal, o qual rende, no conjunto de teste, os seguintes valores:

Tabela 6.2 - Medidas de erro associadas ao modelo de regressão linear com transformação logarítmica, no conjunto de teste

Conjunto	MSE	RMSE	MAE
Teste	5364517	2316.143	877.1341

Este modelo parece ter um comportamento similar ou equivalente nos conjuntos de treino, validação e teste (tal como o das frequências).

Porém, em nenhum dos casos o total de prémios puros chega sequer a cobrir o montante das perdas agregadas devidas a sinistros não-catastróficos, e isto deve-se ao facto de estarmos a recorrer a um valor esperado (ou, melhor dito, a um produto de valores esperados), em vez de quantis, os quais nos oferecem maiores garantias de resiliência face a perdas.

Tabela 6.3 – Comparação entre os custos base totais verificados com sinistros e os prémios puros resultantes do produto das frequências esperadas (modeladas através de uma regressão de Poisson) com as severidades esperadas (modeladas de duas formas distintas)

Custos base totais	Prémios puros totais (modelo log-Normal)	Prémios puros totais (modelo Gama)
1880374	1751955	1867835

Por outro lado, o número total de sinistros previstos pelo modelo de regressão de Poisson coincide exatamente com o número total de sinistros efetivamente observado neste conjunto. A tendência para sub-estimar os prémios atuarialmente necessários e justos tende então a ocorrer nos modelos de custos, sobretudo no modelo log-Normal. Já no modelo Gama, os prémios são desadequados, mas por menor diferença, sendo que este modelo estima corretamente os custos com sinistros, o que nos leva a pensar que, apesar de ainda

válido, o pressuposto de independência entre as variáveis frequência e severidade não se verifica na totalidade.

Tabela 6.4 - Comparação entre os custos base totais verificados com sinistros e os custos totais previstos pelos dois modelos considerados para as severidades esperadas

Custos base totais	Custos totais previstos (modelo log-Normal)	Custos totais previstos (modelo Gama)
1880374	1766064	1880374

Por estes motivos, será necessária a aplicação de uma margem de segurança. Mas qual será a abordagem a seguir na definição de prémios - valor esperado ou quantis? E caso seja aplicado o princípio do valor esperado, qual é a margem de segurança a impor?

Resulta que estas perguntas se respondem uma à outra. Para o cálculo de prémios, iremos usar o princípio do valor esperado, de modo a aproveitar os fatores multiplicativos que as regressões nos dão, e que estão diretamente associadas aos valores médios das frequências e dos custos. No entanto, para saber qual a margem de segurança a aplicar neste princípio, iremos gerar via simulação/reamostragem a distribuição das perdas totais esperadas, de maneira a obter um dado *value-at-risk* o qual, enquanto quantil, dir-nos-á qual é o montante global que serve para cobrir na totalidade as perdas esperadas em, digamos, 95% dos casos.

```
simul <- read.csv2("./pipeline/3_modelos/1_entrada/simul.csv")
```

Iremos considerar 20000 reamostras com 20000 observações, e depois converter, por mera proporcionalidade, os montantes obtidos nestas 20000 observações para a dimensão do conjunto de treino de frequências.

```
# Média de perdas totais por apólice (um elemento do vetor por reamostra)

num.reamostras <- 20000
num.elementos.por.reamostra <- 20000

medias <- rep(NA, num.reamostras)

for(i in 1:num.reamostras){

  medias[i] <- mean(sample(simul$perdas_totais, num.elementos.por.reamostra, replace = TRUE))

}

quantile(medias * nrow(frequencias_treino), probs = c(0.5, 0.9, 0.95, 0.99, 0.995, 0.9995))
```

Tabela 6.5 – Quantis obtidos, via reamostragem e para diferentes probabilidades, das perdas globais registadas no conjunto de treino

Probabilidade associada	0.5	0.9	0.95	0.99	0.995	0.9995

Quantil obtido (em €)	1883642	2213827	2326077	2547271	2647565	2951761

Vamos supôr que o VaR do nosso interesse é o de 95%. Nesse caso, precisamos de receber prémios puros no valor total de 2326077 €, mas os prémios puros com base numa distribuição log-normal constituem um total de apenas 1751955 €, pelo que todas as apólices devem ver os seus prémios estimados ser multiplicados por um fator de 1.3277. efetivamente aplicando nos mesmos uma margem de segurança entre 30% e 35%, podendo esta multiplicação ser implementada diretamente na apólice-padrão, dado que os prémios de todas as outras apólices dependem desta.

Interpretação do modelo obtido e geração de estimativas/previsões

Olhando para os *outputs* fornecidos pelo R, percebemos que os estrutura tarifária a considerar para a modelação de prémios puros é a seguinte:

Tabela 6.6 – Estrutura tarifária obtida, por combinação dos modelos escolhidos para a frequência e para a severidade de sinistros

Fator tarifário	Classes	Fator (multiplicativo)
Nenhum (apólice base)	Ver em baixo	3.98 €
Fator/variável 1	Categoria A (nível base)	× 1.000
	Categoria B	× 1.997
	Categoria C	× 0.600
	Categoria D	× 2.318
	Categoria E	× 1.964
	Categoria F	× 1.929
	Categoria G	× 1.927
	Categoria H	× 0.497
	Categoria I	× 0.505
	Categoria J	× 2.436
	Categoria K	× 1.728
	Categoria L	× 1.660
	Categoria M	× 2.036
	Categoria N	× 0.983

	Categoria O	× 0.870
	Categoria P	× 1.242
	Categoria Q	× 1.038
	Categoria R	× 2.983
	Categoria S	× 2.206
	Categoria T	× 2.112
Fator/variável 2	Categoria A (nível base)	× 1.000
	Categoria B	× 2.120
	Categoria C	× 2.959
	Categoria D	× 3.497
	Categoria E	× 4.080
	Categoria F	× 7.093
Fator/variável 3	Categoria A (nível base)	× 1.000
	Categoria B	× 1.110
Fator/variável 5	Categoria A (nível base)	× 1.000
	Categoria B	× 1.297
	Categoria C	× 1.194
Fator/variável 6	Categoria A	× 0.352
	Categoria B (nível base)	× 1.000
Fator/variável 7	Categoria A (nível base)	× 1.000
	Categoria B	× 1.086

A apólice-base diz respeito a um imóvel com a categoria A (nível base) em cada um dos fatores tarifários utilizados.

Conclusões

O objetivo deste trabalho era o de, a partir de conjuntos de dados históricos, construir uma estrutura tarifária, isto é, uma tabela que permitisse aplicar (se necessário) fatores multiplicativos a uma apólice base ou padrão, de maneira a se conseguir precificar toda e qualquer apólice possível.

Assim, este trabalho serviu para ilustrar a relevância dos modelos lineares generalizados na área atuarial, onde estes são muito populares. Para além disso, serviu para mostrar ao autor deste documento que a aplicação dos mesmos não é tão direta quanto inicialmente seria de esperar, uma vez que apesar de serem vistos como extensões ao modelo clássico de regressão linear, a verdade é que o diagnóstico dos mesmos é mais difícil, perdendo-se ainda algumas propriedades muito interessantes das quais a regressão linear goza, sobretudo ao nível dos estimadores de mínimos quadrados (cuja variância não depende dos próprios valores dos coeficientes, ao contrário do que sucede nos modelos lineares generalizados).

Aqui fica, contudo, patente a dualidade existente entre o modelo regressão linear e os modelos lineares generalizados: se os primeiros têm a vantagem de serem mais facilmente entendidos na sua totalidade, a verdade é que também fazem uso de pressupostos inicialmente algo restritivos. Por exemplo, nem frequências nem severidades de sinistros se prestam a uma modelação por via da distribuição Normal - pelo menos, não sem o recurso a transformações prévias - dado estarem sempre associados a valores não negativos (sendo até discretos no primeiro caso, pois resultam de contagens).

Assim, este trabalho serviu para entender que existe um compromisso: no geral, quanto mais simplista for um pressuposto (ou conjunto de pressupostos), mais fácil se torna a interpretação do modelo com base nele construído mas, em contrapartida, passa a ser mais difícil este pressuposto ser verificado na prática. Apesar disto, verificou-se que os pressupostos do modelo de regressão linear são fáceis de verificar na prática, mesmo que não de uma forma direta e exata (através, por exemplo, de transformações como a logarítmica, aqui utilizada, e também através da teoria assintótica subjacente ao modelo de regressão linear, mais concretamente, subjacente aos seus coeficientes).

Porém, todas estas considerações sobre modelos só significam algo se o recurso aos mesmos fizer sentido. Para avaliar se tal é verdade, teremos de percorrer de forma breve os temas abordados neste trabalho. Foram apresentados tópicos cuja importância não deve ser descurada, tais como:

- A apresentação da ASP (e de entidades associadas), enquanto companhia de seguros;
- Os diferentes ramos e tipos de seguro (com destaque para os comercializados pela ASP);
- Os benefícios económicos e sociais associados à atividade seguradora;
- A legislação e regulação aplicável/em vigor;
- A relevância do atuário enquanto profissional;
- O impacto do contexto económico e social, sobretudo ao nível de baixas taxas de juro e ao nível da pandemia causada pelo covid-19.

Foi também apresentada a teoria económica associada à atividade seguradora, nomeadamente ao nível de conceitos como:

- A existência de um ciclo de negócio invertido - companhias de seguros cobram prémios antes de prestar benefícios contratualmente previstos e, como tal, devem garantir que cobram montantes suficientes para facer face às suas responsabilidades, o que é difícil, porque os custos associados a estas responsabilidades são desconhecidos à *priori* e, portanto, aleatórios, sendo no entanto estimados com maior precisão à medida que se conhece o histórico de mais apólices, pela Lei dos Grandes Números;
- A mutualidade e a solidariedade - encaixando-se a atividade da ASP no primeiro termo, concluimos que os seus seguros devem ser precificados com base nas caraterísticas dos riscos aceites, e não com base nos rendimentos de quem procura protecção;
- Risco *vs* incerteza - a existência de riscos (seguráveis) a cobrir pela ASP implica a representação dos mesmos através de variáveis aleatórias, com distribuições de probabilidade associadas, as quais nos serão muito úteis e sem as quais haveria lugar à incerteza, cenário no qual a precificação de apólices seria (quase) impossível;
- Seleção adversa - mais importante (para este trabalho) do que o risco moral, consiste basicamente na ideia de que maiores riscos devem estar sujeitos a maiores prémios.

A leitura sequencial de todos estes pontos permite-nos identificar as nossas necessidades. Precisamos então de encontrar técnicas que nos permitam estimar prémios que sejam suficientes, distinguindo riscos de acordo com as suas caraterísticas. Por isso, recorremos a modelos de regressão, dado que estes servem precisamente para prever valores de variáveis quantitativas (opondo-se, por isso, a problemas de classificação), com base em variáveis explicativas, variáveis preditivas ou covariáveis. Estes modelos de regressão permitem estimar parâmetros como o valor médio e a variância de cada variável resposta Y_i, servindo efetivamente para ajustar uma distribuição a cada risco individual, o que é precisamente o que pretendemos para evitar fenómenos de seleção adversa. Ao terem natureza probabilística (por estarmos interessados em mais do que um mero ajuste de curvas), a escolha de modelos lineares (generalizados) foi natural e está, portanto, justificada. Esta escolha motiva então a apresentação de teoria antes da modelação propriamente dita, sendo esta teoria apresentada de uma forma mais geral e não tão aprofundada.

Verificamos, então, que as justificações para recorrer a modelos de regressão na construção de uma estrutura tarifária são, sobretudo, de índole económica; por isso, a escolha do tipo de modelos segue uma abordagem holística. Esta preocupação verifica-se também noutras fases; em particular, o modelo inicial é construído de forma a conter todas as variáveis preditivas consideradas pertinentes (sobretudo depois de uma análise exploratória aos dados), procedendo-se de seguida à remoção sequencial de toda e qualquer variável cuja exclusão seja estatisticamente recomendada (através do AIC e/ou de algum *t-test* individual), se tais variáveis existirem e se não houverem bons motivos lógicos ou de negócio que impeçam esta remoção. De notar ainda que, não sendo elevado o número de covariáveis com as quais estamos a trabalhar (tendo em mente que é atribuída uma variável binária a cada valor de uma covariável categórica), procedeu-se à uma construção manual e não automatizada de modelos, de maneira a ter maior controlo neste processo e ser capaz de explicá-lo.

Também a validação de modelos seguiu uma abordagem mais abrangente. Como já vimos, a exigência de uma abordagem probabilística fez com que optassemos por um caminho mais tradicional, como o dos modelos lineares generalizados (em detrimento, por exemplo, de árvores de regressão), pelo que o tipo de modelos escolhido parece adequado. O modelo clássico de regressão linear com transformação

logarítmica para a severidade foi, dos três considerados, o mais fácil de diagnosticar, dada a sua natureza mais simplista e o facto de ter sido abordado em disciplinas do mestrado. Já em relação aos modelos de regressão de Poisson (para frequências) e de regressão Gama (para severidades), este diagnóstico tornou-se mais difícil e menos claro. Daí uma das grandes vantagens de uma abordagem mais diversa: graças à separação de observações em conjuntos de treino, validação e teste, foi possível "conquistar" mais um método de validação. Também foi útil a obtenção e elaboração de gráficos caraterísticos da prática atuarial, com dois eixos verticais - um para a frequência ou severidade associada a apólices ou sinistros (à direita), e outro para a frequência (absoluta ou relativa) da variável categórica que serve para a obtenção de indicadores condicionais.

Dado tudo o que já foi dito, podemos afirmar que os modelos estatísticos obtidos são de qualidade e que se ajustam razoavelmente bem aos dados, dadas as caraterísticas dos fenómenos a modelar. Isto verifica-se sobretudo na modelação da frequência de sinistros por apólice, uma vez que o número de sinistros é bem inferior ao número de apólices e, como tal, é mais difícil obter modelos com o mesmo nível de ajuste para a severidade. Estes modelos são ainda interpretáveis (ao contrário do que acontece, por exemplo, com redes neurais), podendo o funcionamento dos mesmos ser explicado de uma forma intuitiva a tomadores de decisão, entidades de supervisão e demais *stakeholders*.

Podemos concluir que a elaboração deste trabalho resultou, para além do próprio documento:

- Na obtenção de uma estrutura tarifária devidamente fundamentada, a qual representa uma melhoria face à estrutura tarifária atualmente em vigor na ASP;
- No desenvolvimento de conhecimentos em diversas áreas, com especial destaque naturalmente para os modelos lineares generalizados;
- Na obtenção de experiência profissional;
- Numa metodologia integrada para a elaboração de outros projetos de análise de dados;
- No desenvolvimento de aptidões de escrita de documentos com linguagem fluida e correta.

Ainda relativamente à análise de dados, algumas sugestões para melhores resultados passam por:

- Implementar, em projetos futuros, a metodologia integrada já considerada numa fase inicial;
- Procurar saber mais sobre a qualidade, quantidade e variedade de dados que é possível obter;
- Por refletir regularmente sobre os objetivos da análise/projeto e prestar maior atenção, na análise exploratória de dados, às relações de maior (potencial) interesse, não explorando exaustivamente todas as relações possíveis.

Ao nível de aprendizagens futuras, podemos identificar vários caminhos a explorar, em vertentes:

- Quantitativas - tanto ao nível de conceitos estatísticos, como ao nível da teoria matemática que lhes serve de base;
- Computacionais - tanto ao nível do cumprimento de boas práticas de programação no geral, como ao nível do uso das ferramentas presentes em linguagens como o R;
- De área de aplicação - tanto ao nível de aspetos económicos e de gestão comuns a todas as organizações e empresas, como ao nível do funcionamento específico do setor segurador.

Bibliografia consultada

Aegon Santander Portugal. (s.d.). Obtido de Aegon Santander Portugal: https://www.aegon-santander.pt/

Associação Portuguesa de Seguradores. (s.d.). Obtido de Associação Portuguesa de Seguradores: https://www.apseguradores.pt/pt/

Autoridade de Supervisão de Seguros e Fundos de Pensões. (s.d.). Obtido de Autoridade de Supervisão de Seguros e Fundos de Pensões: https://www.asf.com.pt/

Bahnemann, D. (s.d.). *Distributions for actuaries.* Obtido de Casualty Actuarial Society: https://www.casact.org/pubs/monographs/papers/02-Bahnemann.pdf

Como funciona o seguro. (s.d.). Obtido de Insurance Europe: https://www.insuranceeurope.eu/sites/default/files/attachments/How%20insurance%20works%20-%20Portuguese%20translation.pdf

European insurers and the curse of low interest rates. (s.d.). Obtido de The Economist: https://www.economist.com/finance-and-economics/2016/12/15/european-insurers-and-the-curse-of-low-interest-rates

Fox, J. (s.d.). Generalized Linear Models. Em J. Fox, *Applied Regression & Generalized Linear Models.* SAGE Publications. Obtido de Generalized Linear Models: https://kilpatrick.eeb.ucsc.edu/wp-content/uploads/2015/04/GLMs-Chapter_15.pdf

Goldburd, M., Khare, A., Tevet, D., & Guller, D. (s.d.). *Generalized linear models for insurance rating.* Obtido de Casualty Actuarial Society: https://www.casact.org/pubs/monographs/papers/05-Goldburd-Khare-Tevet.pdf

Insurance and coronavirus (Covid-19): our expectations of firms. (s.d.). Obtido de Financial Conduct Authority: https://www.fca.org.uk/firms/insurance-and-coronavirus-our-expectations

Insurance and economic growth. (s.d.). Obtido de Generali: https://www.generali.com/info/discovering-generali/all/2018/Insurance-and-economic-growth

Kim, B. (s.d.). *Understanding Diagnostic Plots for Linear Regression Analysis.* Obtido de University of Virginia Library - Research Data Services + Sciences: https://data.library.virginia.edu/

McCullagh, P., & Nelder, J. (1983). *Generalized Linear Models.* London: Chapman and Hall.

Portugués, E. G. (s.d.). *Generalized linear models.* Obtido de Notes for Predictive Modeling: https://bookdown.org/egarpor/PM-UC3M/glm.html

Rego, M. L., & Silva, R. C. (s.d.). *Catástrofes naturais e seguros.* Obtido de Repositório da Universidade Nova de Lisboa: https://run.unl.pt/bitstream/10362/15130/1/Cat%c3%a1strofes%20Naturais%20e%20Seguros%20-%20Margarida%20Lima%20Rego.pdf

Rodríguez, G. (s.d.). *Generalized Linear Models*. Obtido de Princeton University: https://data.princeton.edu/wws509/notes/

What are pseudo R-Squareds? (s.d.). Obtido de UCLA Institute for Digital Research & Education - Statistical Consulting: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/

Apontamentos da professora Teresa Alpuim para apoio à disciplina de Modelos Lineares do Mestrado em Matemática Aplicada à Economia e Gestão.

Apontamentos da professora Marília Antunes para apoio à disciplina de Análise da Variância e Regressão do Mestrado em Matemática Aplicada à Economia e Gestão.

Apontamentos da professora Marli Amorim Ferreira para apoio à disciplina de Atividade Seguradora do Mestrado em Matemática Aplicada à Economia e Gestão.

Sebenta da professora Gracinda Rita Guerreiro para apoio à disciplina de Gestão do Risco em Actuariado Não Vida do curso do Mestrado em Actuariado, Estatística e Investigação Operacional, da Faculdade de Ciências e Tecnologia da Universidade Noval de Lisboa.

Anexos

Hierarquia e estrutura do projeto de tarifação

De uma forma específica, e por ordem de relevância, as pastas a considerar no projeto de tarifação serão as seguintes:

- **pipeline** - Refletem o percurso de exploração seguido pelo autor da análise, através de *notebooks* e ficheiros associados:
    - **1_dados** - Dados em vários formatos (por exemplo, Excel, CSV ...):
    - **2_aed** - Análise exploratória de dados:
    - **3_modelos** - Modelos treinados/ajustados, previsões e demais elementos:
- **resultados** - Resultados finais das análises nos mais diversos formatos:
    - **apresentacoes** - Versões finais de apresentações baseadas em diapositivos (por exemplo, ficheiros .pptx) obtidas mediante exportação de ficheiros em RMarkdown, e posteriormente editadas no MS Office;
    - **graficos** - Gráficos obtidos através de análises (e que tenham especial relevância);
    - **interactivos** - Resultados interativos (por exemplo, *dashboards*);
    - **modelos** - Modelos estatísticos construídos;
    - **outros** - Todo e qualquer elemento que não pertença a outro subdiretório;
    - **relatorios** - Versões finais de documentos/relatórios (por exemplo, ficheiros .docx ou .pdf) obtidos mediante exportação de ficheiros em RMarkdown, e posteriormente editados no MS Office;
- **scripts** - Todo e qualquer código a executar (por exemplo, num servidor), organizado em ficheiros ordenados por números no início do nome dos mesmos, de maneira a indicar a ordem de execução:
    - **1_dados** - *Scripts* que geram ou descarregam conjuntos de dados e transformam-nos em conjuntos prontos/aptos para fases futuras (exploração e, posteriormente, modelação);
    - **2_aed** - *Scripts* que geram visualizações exploratórias, estatísticas de sumário, tabelas de frequências, entre outros indicadores (AED - análise exploratória de dados);
    - **3_modelos** - *Scripts* que treinam/ajustam modelos com o intuito de realizar inferências e previsões;
- **referencias** - Referências para o projeto (ao nível de conhecimento):
    - **links** - Hiperligações para conteúdos relevantes na Internet;
    - **outros** - Todo e qualquer elemento que não pertença a outro subdiretório;
    - **papers** - *Papers* académicos;
- **misc** - Miscelânea ("tudo o resto"):
    - **cache** – Cache;
    - **configuracoes** - Ficheiros de configuração;
    - **documentos** - Documentação criada (por exemplo, de suporte);
    - **ficheiros_temporarios** - Ficheiros com caráter temporário;
    - **outros** - Todo e qualquer elemento que não pertença a outro subdiretório;
    - **rascunhos** - Rascunhos criados;
    - **registos** - Registos (*logs*) de execução de códigos;
    - **scripts** - Código fonte para ferramentas descarregadas ou de outra forma utilizadas no decorrer da análise (e não geradas pela mesma);
    - **testes** - Testes efetuados ao código-fonte;

- **pacotes** - Pacotes (extensões) descarregados para uso em diversos ambientes;
- **ajuda.Rmd** - Ficheiro de referência para obtenção de ajuda, o qual deve incluir também os requisitos para reproduzir o ambiente no qual a análise decorreu;
- **ajuda.html** - Resultado da exportação de ajuda.Rmd;
- **ajuda.doc** - Resultado da exportação de ajuda.Rmd;
- <**outros ficheiros e atalhos**>.

Em maior detalhe, temos a seguinte distinção:

- *Scripts* são meros automatismos que permitem transformar imediatamente e de forma sucessiva *inputs* (como dados) nos *outputs* finais (como modelos), sem grandes explicações, permitindo assim a passagem a produção/execução em servidores;
- Resultados são elementos prontos para apresentação, com mais explicações (as quais, excetuando comentários, não estão presentes em *scripts*);
- Um *pipeline* é, neste contexto, um processo necessário, do qual podem resultar tanto *scripts* como resultados prontos para apresentação, e que conta a história de uma exploração de raciocínios, construção de conhecimentos e obtenção de informações.

O seguimento das práticas aqui descritas tem a vantagem de levar à existência de vários pontos de segurança, para que seja possível identificar e corrigir quaisquer erros de forma célere. Adicionalmente:

- Pastas que não venham a ser usadas podem ser removidas;
- Algumas pastas terão uma numeração associada correspondente à sua ordem de execução;
- O *working directory* é definido, através do R Projects, como sendo a pasta do projeto do mais elevado nível, prática que nos permitirá usar *relative paths* para nos referirmos aos diversos diretórios desta estrutura e tornar o nosso trabalho portátil. Se usarmos *relative paths* e o *working directory* for o especificado, podemos copiar a pasta do nosso projeto para outro computador e o código presente funcionará sem nenhum problema.

Scripts criados pelo autor

Em **./misc/scripts** podemos encontrar três scripts – **dpp.R**, **eda.R** e **mod.R** – cujos conteúdos são, respetivamente, apresentados de seguida:

```
# Script dpp.R

dpp.prop.obs <- function(data, cond, freqs = "rel", digits = 3){

 if(freqs == "rel"){

   tmp.1 <- length(data[!is.na(data) == TRUE])
   tmp.2 <- length(data[!is.na(data) == TRUE & cond])
   result <- round(tmp.2/tmp.1, digits)

 } else if(freqs == "abs") {
```

```
   result <- length(data[!is.na(data) == TRUE & cond])

  }

  return(result)

}

# dim(dataset)
# names(dataset)
# class(dataset)

dpp.data.anomalies <- function(data, freqs = "rel", digits = 3){

  if(freqs == "rel"){

    tmp.1 <- length(data)
    tmp.2 <- length(data[is.na(data) | is.infinite(data) | is.nan(data) | is.null(data)])
    result <- round(tmp.2/tmp.1, digits)

  } else if(freqs == "abs") {

    result <- length(data[is.na(data) | is.infinite(data) | is.nan(data) | is.null(data)])

  }

  return(result)

}

dpp.train.test.split <- function(df, p_train_1, p_train_2 = 1){

  n <- nrow(df)

  s_train <- floor(p_train_1 * n)

  train_ind <- sample(1:n, s_train, replace = FALSE)

  train_train_ind <- sample(train_ind,
                            floor(p_train_2 * length(train_ind)),
                            replace = FALSE)
  train_val_ind <- setdiff(train_ind, train_train_ind)
  test_ind <- setdiff(1:n, train_ind)

  result <- list(c(train_train_ind), c(train_val_ind), c(test_ind))

  names(result) <- c("train.ind", "val.ind", "test.ind")

  return(result)

}
```

```
dpp.recode.exact <- function(data, orig, new){
  result <- data
  for (i in 1:length(orig)){
    result <- replace(result, data == orig[i], new[i])
  }
  return(result)
}
dpp.recode.interval <- function(data, orig, new){
  result <- data
  for(i in 1:(length(orig) - 1)){
    result[data >= orig[i] & data < orig[i+1]] <- new[i]
  }
  return(result)
}
dpp.rmd.table.display <- function(x, mode = "HTML"){
  if(mode == "word" | mode == "Word"){
    result <- kable(x)
  } else {
    result <- x
  }
  return(result)
}
dpp.rmd.table.display <- function(x, mode = "HTML"){
  if(mode%in%c("word", "Word", "powerpoint", "PowerPoint", "office", "Office")){
    result <- kable(x)
```

```
 } else if(mode%in%c("html", "HTML", "web", "Web")) {

  result <- x

 }

 return(result)

}
```

```
# Script eda.R

color.palette <- c("#F00E0E", "#F54B42", "#F58014", "#FFDD00",
                   "#86FF24", "#19FF4F", "#24E3BA", "#38DEFF",
                   "#2166FC", "#7433FF", "#AF19D1", "#F52C9A")

red.palette <- c("#ff0000", "#ff1f1f", "#ff3f3f", "#ff5f5f",
                 "#ff7f7f", "#ff9f9f", "#ffbfbf", "#ffdfdf", "#ffffff")

orange.palette <- c("#ff8000", "#ff8f1f", "#ff9f3f", "#ffaf5e", "#ffc080",
                    "#ffd0a0", "#ffdfc0", "#ffefe0", "#ffffff")

yellow.palette <- c("#ffff00", "#ffff1f", "#ffff3f", "#ffff5f",
                    "#ffff7f", "#ffff9f", "#ffffbf", "#ffffdf", "#ffffff")

green.palette <- c("#40ff00", "#57ff1f", "#70ff40", "#8aff5f",
                   "#a2ff7f", "#b9ffa0", "#d0ffbf", "#e9ffe0", "#ffffff")

blue.palette.light <- c("#00ffff", "#1fffff", "#3fffff", "#5fffff",
                        "#7fffff", "#9fffff", "#bfffff", "#dfffff", "#ffffff")

blue.palette.dark <- c("#0000ff", "#1f1fff", "#3f3fff", "#5f5fff",
                       "#7f7fff", "#9f9fff", "#bfbfff", "#dedeff", "#ffffff")

purple.palette <- c("#bf00ff", "#c71fff", "#cf40ff", "#d75eff",
                    "#df80ff", "#e79eff", "#efbfff", "#f7deff", "#ffffff")

pink.palette <- c("#ff00a0", "#ff1fab", "#ff40b7", "#ff61c3", "#ff80cf",
                  "#ff9eda", "#ffbfe7", "#ffdef3", "#ffffff")

black.palette <- c("#000000", "#1f1f1f", "#404040", "#5e5e5e", "#808080",
                   "#a1a1a1", "#bfbfbf", "#e0e0e0", "#ffffff")

eda.prop.obs <- function(data, cond, freqs = "rel", digits = 3){

 if(freqs == "rel"){

   tmp.1 <- length(data[!is.na(data) == TRUE])
   tmp.2 <- length(data[!is.na(data) == TRUE & cond])
   result <- round(tmp.2/tmp.1, digits)
```

```
  } else if(freqs == "abs") {

    result <- length(data[!is.na(data) == TRUE & cond])

  }

  return(result)

}

# dim(dataset)
# names(dataset)
# class(dataset)

eda.data.anomalies <- function(data, freqs = "rel", digits = 3){

  if(freqs == "rel"){

    tmp.1 <- length(data)
    tmp.2 <- length(data[is.na(data) | is.infinite(data) | is.nan(data) | is.null(data)])
    result <- round(tmp.2/tmp.1, digits)

  } else if(freqs == "abs") {

    result <- length(data[is.na(data) | is.infinite(data) | is.nan(data) | is.null(data)])

  }

  return(result)

}

eda.skewness <- function(data, na.rm = TRUE, digits = 3){

  centered.data <- data - mean(data, na.rm = na.rm)

  result <- (mean(centered.data ^ 3, na.rm = na.rm)) / ((mean(centered.data ^ 2, na.rm = na.rm)) ^ 1.5)

  result <- round(result, digits)

  return(result)

}

eda.kurtosis <- function(data, na.rm = TRUE, digits = 3){

  centered.data <- data - mean(data, na.rm = na.rm)

  result <- (mean(centered.data ^ 4, na.rm = na.rm)) / ((mean(centered.data ^ 2, na.rm = na.rm)) ^ 2)
```

```
  result <- round(result, digits)

  return(result)

}

eda.univariate.numeric.stats <- function(data, desc = NULL, na.rm = TRUE, digits = 3){

  result <- c(round(n <- length(data[!is.na(data) == na.rm])),
              round(mean(data, na.rm = na.rm), digits),
              round(quantile(data, c(0.50), na.rm = na.rm), digits),
              round(as.numeric(names(sort(table(data), decreasing = TRUE))[1]), digits),
              round(var(data, na.rm = na.rm), digits),
              round(sd(data, na.rm = na.rm), digits),
              round(sd(data, na.rm = na.rm)/sqrt(n), digits),
              round(sd(data, na.rm = na.rm)/mean(data, na.rm = na.rm), digits),
              round(quantile(data, c(0.25), na.rm = na.rm), digits),
              round(quantile(data, c(0.75), na.rm = na.rm), digits),
              round(IQR(data, na.rm = na.rm), digits),
              round(min(data, na.rm = na.rm), digits),
              round(max(data, na.rm = na.rm), digits),
              round(max(data, na.rm = na.rm) - min(data, na.rm = na.rm), digits),
              round(eda.skewness(data), digits),
              round(eda.kurtosis(data), digits))

  names(result) <- c("no.obs", "mean", "median", "mode", "variance", "std.deviation", "std.error",
                     "coef.variation", "first.quartile", "third.quartile", "inter.quartile.range", "minimum",
                     "maximum", "abs.range", "skewness", "kurtosis")

  result <- as.data.frame(result)

  colnames(result) <- desc

  return(result)

}

# eda.univariate.numeric.stats.batch <- apply(dataset, 2, eda.univariate.numeric.stats, desc = NULL, na.r
m = na.rm)

eda.joint.numeric.stats <- function(data, na.rm = TRUE, collin = FALSE, threshold = 0.75, digits = 3){

  result <- list(no.obs = nrow(data),
                 no.variables = ncol(data),
                 mean.avg = round(colMeans(data, na.rm = na.rm), digits),
                 covariance = round(cov(data), digits),
                 correlation = round(cor(data), digits))

  if(collin == TRUE){

    result$removal <- (cor(data) >= threshold)
```

```
  }

  return(result)

}

eda.conditional.numeric.stats <- function(data, cond, na.rm = TRUE, digits = 3, comparison = FALSE,
                                  col.names = c("Conditional", "Not conditional", "Comparison")){

  result <- do.call("eda.univariate.numeric.stats", list(data[cond], digits = digits, na.rm = na.rm))

  if(comparison == TRUE){

    result <- cbind(result, eda.univariate.numeric.stats(data, digits = digits, na.rm = na.rm))

    result <- data.frame(result, ifelse(result[, 1] > result[, 2], "Bigger",
                                 ifelse(result[, 1] == result[, 2], "Equal", "Smaller")))

    colnames(result) <- col.names

  }

  return(result)

}

eda.conditional.numeric.groupby.stats <- function(data, cond.factor,
                                          main = "", sub = "", xlab = "",
                                          ylab = "", na.rm = TRUE){

  result <- cbind(aggregate(data, by = list(group.by = cond.factor), length),
                  aggregate(data, by = list(group.by = cond.factor), function(x) mean(x, na.rm = na.rm))[, 2],
                  aggregate(data, by = list(group.by = cond.factor), function(x) var(x, na.rm = na.rm))[, 2],
                  aggregate(data, by = list(group.by = cond.factor), function(x) sd(x, na.rm = na.rm))[, 2],
                  aggregate(data, by = list(group.by = cond.factor), function(x) quantile(x, na.rm = na.rm))[, 2]
,
                  aggregate(data, by = list(group.by = cond.factor), function(x) IQR(x, na.rm = na.rm))[, 2],
                  aggregate(data, by = list(group.by = cond.factor), function(x) range(x, na.rm = na.rm))[, 2],
                  aggregate(data, by = list(group.by = cond.factor), function(x) eda.skewness(x, na.rm = na.rm
))[, 2],
                  aggregate(data, by = list(group.by = cond.factor), function(x) eda.kurtosis(x, na.rm = na.rm))
[, 2])

  result <- cbind(result, as.numeric(result[, 5]) / sqrt(as.numeric(result[, 2])), as.numeric(result[, 5]) / as
.numeric(result[, 3]))

  result <- result[, -13]

  names(result) <- c("factor/group", "no.obs", "mean", "variance", "std.deviation",
                     "minimum", "first.quartile", "median", "third.quartile",
```

```
                "maximum", "inter.quartile.range", "abs.range", "skewness", "kurtosis",
                "std.error", "coef.variation")

  return(result)

}

eda.conditional.numeric.groupby.plots <- function(data, cond.factor, k,
                                    main = "", sub = "", xlab = "",
                                    ylab = "", ylim = NULL, box.col = "#71BD00"){

  boxplot(data ~ cond.factor, main = main, sub = sub, xlab = xlab, ylab = ylab,
        ylim = ylim, col = box.col, range = k)

}

# eda.data.anomalies.batch <- apply(dataset, 2, eda.data.anomalies, freqs = freqs)

eda.joint.numeric.plots <- function(dataset, dot.col = "#00AFBB",
                            line.col = "#CF1717", lwd = 2){

  # Correlation panel

  panel.cor <- function(x, y){

    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- round(cor(x, y), digits = 2)
    txt <- paste0("R = ", r)
    cex.cor <- 0.8
    text(0.5, 0.5, txt, cex = 3 * cex.cor * sqrt(abs(r)))

  }

  # Customize upper panel

  upper.panel <- function(x, y, pch = 19){

    points(x, y, pch = pch, col = dot.col)
    abline(lm(y ~ x), col = line.col, lwd = lwd)

  }

  # Create the plots

  pairs(dataset, lower.panel = panel.cor, upper.panel = upper.panel)

}

eda.univariate.numeric.plots <- function(data, main = "", sub = "", xlab = "",
                                ylab = "", xlim = NULL, ylim = NULL,
```

```
                                bars.col = color.palette[9],
                                line.col = color.palette[1], lwd = 2, breaks = NULL){

  if(is.integer(data[!is.na(data)]) == TRUE | is.numeric(data[!is.na(data)]) == TRUE){

    hist(data[!is.na(data)], breaks = breaks, freq = FALSE, col = bars.col, main = main, xlab = xlab, ylab =
ylab, xlim = xlim, ylim = ylim)
    lines(density(data[!is.na(data)]), col = line.col, lwd = lwd)

  } else {

    warning("The variable entered is not quantitative!")

  }

}

eda.univariate.categorical.plots <- function(data, main = "", sub = "", xlab = "",
                                    ylab = "", xlim = NULL, ylim = NULL,
                                    bars.freqs = "rel", bars.col = color.palette,
                                    sort = TRUE, no.elem = NULL,
                                    names = NULL, agg.name = "Other"){

  if(is.integer(data) == TRUE | is.factor(data) == TRUE | is.character(data) == TRUE | is.logical(data) =
= TRUE){

    result <- table(data[!is.na(data)])

    if(bars.freqs == "rel"){

      result <- prop.table(result)

    }

    if(sort == TRUE){

      result <- sort(result, decreasing = TRUE)

    }

    if(!is.null(no.elem)){

      tmp <- rep(NA, no.elem + 1)
      tmp[1:no.elem] <- result[1:no.elem]
      names(tmp)[1:no.elem] <- names(result)[1:no.elem]

      tmp[no.elem + 1] <- sum(result[no.elem + 1:length(result)][1:(length(result) - no.elem)])
      names(tmp)[no.elem + 1] <- agg.name

      result <- tmp
```

```
  }

  if(is.null(names)){

    desc <- names(result)

  }

  barplot(result, col = bars.col, main = main, sub = sub, xlab = xlab, ylab = ylab, xlim = xlim, ylim = yli
m, names.arg = desc)

 } else {

  warning("The variable entered is not qualitative!")

 }

}

eda.conditional.numeric.boxplots <- function(data, cond,
                            names = c("Conditional", "Non-conditional"),
                            box.col = "#DBAF0F", main = "", sub = "",
                            xlab = "", ylab = "", xlim = NULL, ylim = NULL,
                            k = 0){

 if(is.integer(data) == TRUE | is.numeric(data) == TRUE){

  boxplot(data[cond], data, names = names, col = box.col,
          main = main, sub = sub, xlab = xlab, ylab = ylab,
          xlim = xlim, ylim = ylim, range = k)

 }

}

eda.conditional.numeric.densities <- function(data, cond, legend = TRUE,
                             legend.names = c("Conditional", "Non-conditional"),
                             legend.pos = "topright", lines.col = color.palette[c(9, 1)], main = "",
                             sub = "", xlab = "", ylab = "",
                             xlim = NULL, ylim = NULL){

 if(is.integer(data) == TRUE | is.numeric(data) == TRUE){

  plot(density(data), main = main, sub = sub, xlab = xlab, ylab = ylab,
       xlim = xlim, ylim = ylim, col = lines.col[1])
  lines(density(data[cond]), col = lines.col[2])

 }

 if(legend == TRUE){
```

```
  legend(legend.pos, legend = legend.names, fill = lines.col)

 }

}

eda.conditional.categorical.barplots <- function(data, cond, main = "", sub = "",
                                xlab = "", ylab = "", ylim = c(0, 0.7),
                                bars.col = color.palette[c(9, 1)], legend = TRUE,
                                legend.names = c("Conditional", "Non-conditional"),
                                legend.pos = "topright", sort.bars = "cond",
                                no.elem = NULL, agg.name = "Other"){

 if(is.integer(data) == TRUE | is.factor(data) == TRUE | is.character(data) == TRUE | is.logical(data)){

  global.table <- prop.table(table(data))
  cond.table <- prop.table(table(data[cond]))

  result <- merge(as.data.frame(global.table), as.data.frame(cond.table),
            by.x = "data", by.y = "Var1", all = TRUE)

  result <- subset(result, select = -c(data))

  result[is.na(result)] <- 0

  result <- as.matrix(t(result))

 }

 if(sort.bars == "cond"){

  cond.table <- prop.table(table(data[cond]))
  cond.table <- sort(cond.table, decreasing = TRUE)

  global.table <- prop.table(table(data))
  global.table <- global.table[names(cond.table)]

  result <- rbind(cond.table, global.table)

 } else if(sort.bars == "uncond"){

  global.table <- prop.table(table(data))
  global.table <- sort(global.table, decreasing = TRUE)

  cond.table <- prop.table(table(data[cond]))
  cond.table <- cond.table[names(global.table)]

  result <- rbind(cond.table, global.table)

 }
```

```
  if(!is.null(no.elem)){

    result[1, no.elem + 1] <- sum(result[1, (no.elem + 1):ncol(result)], na.rm = TRUE)
    result[2, no.elem + 1] <- sum(result[2, (no.elem + 1):ncol(result)], na.rm = TRUE)

    result <- result[, 1:(no.elem + 1)]

    colnames(result)[no.elem + 1] <- agg.name

  }

  result[is.na(result)] <- 0

  barplot(result, beside = TRUE, col = bars.col, main = main, sub = sub, xlab = xlab,
          ylab = ylab, ylim = ylim)

  if(legend == TRUE){

    legend(legend.pos, legend = legend.names, fill = bars.col)

  }

  return(result)

}

eda.tukey.outliers <- function(dataset, column, k = 0) {

  inf <- quantile(dataset[[column]], c(0.25), na.rm = T) - k * IQR(dataset[[column]], na.rm = T)
  sup <- quantile(dataset[[column]], c(0.75), na.rm = T) + k * IQR(dataset[[column]], na.rm = T)

  index <- which(dataset[[column]] < inf | dataset[[column]] > sup)

  return(index)

}

eda.outlier.display <- function(dataset, column, k, cent.ref = "mean",
                                sort.mode = "decreasing"){

  result <- dataset[eda.tukey.outliers(dataset, column, k = k), ]

  if (cent.ref == "mean" & sort.mode == "decreasing"){

    result <- result[order(abs(result[[column]] - mean(result[[column]])), decreasing = TRUE), ]

  } else if (cent.ref == "mean" & sort.mode == "increasing") {

    result <- result[order(abs(result[[column]] - mean(result[[column]])), decreasing = FALSE), ]

  } else if (cent.ref == "mean" & sort.mode == "decreasing") {
```

```
  result <- result[order(abs(result[[column]] - mean(result[[column]])), decreasing = TRUE), ]
 } else if (cent.ref == "mean" & sort.mode == "increasing") {
  result <- result[order(abs(result[[column]] - mean(result[[column]])), decreasing = FALSE), ]
 }
 return(result)
}
eda.recode.exact <- function(data, orig, new){
 result <- data
 for (i in 1:length(orig)){
  result <- replace(result, data == orig[i], new[i])
 }
 return(result)
}
eda.recode.interval <- function(data, orig, new){
 result <- data
 for(i in 1:(length(orig) - 1)){
  result[data >= orig[i] & data < orig[i+1]] <- new[i]
 }
 return(result)
}
eda.rmd.table.display <- function(table, mode = "HTML"){
 if(mode%in%c("word", "Word", "powerpoint", "PowerPoint", "office", "Office")){
  result <- kable(table)
 } else if(mode%in%c("html", "HTML", "web", "Web")) {
  result <- table
```

```
  }

  return(result)

}

eda.bars.lines <- function(dataset, desc.bars, desc.lines, desc.stat = "mean",
                           main = "", sub = "", xlab = "", ylab.bars = "",
                           ylab.lines = "", ylim.bars = "", ylim.lines = "",
                           lwd = 2, bars.col = color.palette,
                           line.col = "red", margins = c(5, 5, 2, 5),
                           conf.int = FALSE, conf.level = 0.95,
                           conf.lines.col = "green", space = 0,
                           values = FALSE){

  means <- aggregate(dataset[[desc.lines]],
                     list(Bars = dataset[[desc.bars]]), desc.stat)
  freqs <- table(dataset[[desc.bars]])

  if(desc.stat == "mean" & conf.int == TRUE){

    vars <- aggregate(dataset[[desc.lines]],
                      list(Bars = dataset[[desc.bars]]), "var")
    ns <- aggregate(dataset[[desc.lines]],
                    list(Bars = dataset[[desc.bars]]), "length")
    ses <- vars$x/ns$x

  }

  if(is.character(ylim.bars)) {

    ylim.bars <- c(0, max(table(dataset[[desc.bars]])))

  }

  if(is.character(ylim.lines)) {

    ylim.lines <- c(floor(0.8 * min(means$x)), ceiling(1.2 * max(means$x)))

  }

  par(mar = margins)

  # barplot(freqs, space = space, col = bars.col, ylab = ylab.bars, ylim = ylim.bars)
  mp <- barplot(freqs, space = space, col = bars.col, ylab = ylab.bars,
                ylim = ylim.bars)

  par(new = TRUE)

  with(means, plot(means$x, main = main, sub = sub, xlab = xlab,
                   type = "l", lwd = lwd, xaxt = "n", yaxt = "n",
```

```
            axes = F, xlim = c(min(mp), max(mp) + 1),
            ylim = ylim.lines, col = line.col, ylab = ylab.bars))

  if(desc.stat == "mean" & conf.int == TRUE){

    with(means, lines(mp + 0.5, means$x + qnorm(conf.level/2) * ses,
                col = conf.lines.col, lwd = lwd))
    with(means, lines(mp + 0.5, means$x + qnorm(1 - conf.level/2) * ses,
                col = conf.lines.col, lwd = lwd))

  }

  axis(side = 4)

  mtext(side = 4, line = 2, ylab.lines)

  if(values == TRUE){

    return(mp)

  }

}

eda.frequency.tables <- function(dataset, var.desc = "var1", freqs = "rel",
                          digits = 3, no.elem = NULL, sort = TRUE,
                          names = "", agg.name = "Other"){

  if(freqs == "rel"){

    result <- prop.table(table(dataset[[var.desc]], dnn = names))
    result <- round(result, digits)

    if(sort == TRUE){

      result <- sort(result, decreasing = TRUE)

    }

    if(!is.null(no.elem)){

      tmp <- rep(NA, no.elem + 1)
      tmp[1:no.elem] <- result[1:no.elem]
      names(tmp)[1:no.elem] <- names(result)[1:no.elem]

      tmp[no.elem + 1] <- sum(result[no.elem + 1:length(result)][1:(length(result) - no.elem)])
      names(tmp)[no.elem + 1] <- agg.name

      result <- tmp

    }
```

```
  } else if(freqs == "abs") {

   result <- table(dataset[[var.desc]], dnn = names)

   if(sort == TRUE){

     result <- sort(result, decreasing = TRUE)

   }

   if(!is.null(no.elem)){

     tmp <- rep(NA, no.elem + 1)
     tmp[1:no.elem] <- result[1:no.elem]
     names(tmp)[1:no.elem] <- names(result)[1:no.elem]

     tmp[no.elem + 1] <- sum(result[no.elem + 1:length(result)][1:(length(result) - no.elem)])
     names(tmp)[no.elem + 1] <- agg.name

     result <- tmp

   }

  }

  return(result)

}

eda.categorical.crosstabs <- function(dataset, vars.desc = c("var1", "var2"),
                                      freqs = "comp", digits = 3, names = "",
                                      lower.lim = 0.8, upper.lim = 1.2,
                                      sort.mode = "expected", expected.min = 5){

  if(freqs == "rel"){

   result <- prop.table(table(dataset[[vars.desc[1]]], dataset[[vars.desc[2]]], dnn = names))
   result <- round(result, digits)


  } else if(freqs == "abs"){

   result <- table(dataset[[vars.desc[1]]], dataset[[vars.desc[2]]], dnn = names)

  } else if(freqs == "comp"){

   observed <- table(dataset[[vars.desc[1]]], dataset[[vars.desc[2]]], dnn = names)
   expected <- table(dataset[[vars.desc[1]]], dataset[[vars.desc[2]]], dnn = names)
   total <- sum(observed)
```

```
  tmp <- cbind(rowSums(expected)) %*% colSums(expected) / total

  for(i in 1:nrow(expected)){

    for(j in 1:ncol(expected)){

      expected[i, j] <- tmp[i, j]

    }

  }

  observed <- as.data.frame(observed)
  expected <- as.data.frame(as.table(expected))

  result <- cbind(observed, expected)
  result <- result[, c(1:3, 6)]

  result[, 4] <- round(result[, 4], digits)

  result <- result[result[, 4] > expected.min, ]

  colnames(result) <- c(vars.desc, "observed", "expected")

  result$ratio <- result$observed / result$expected

  result <- result[result$ratio < lower.lim | result$ratio > upper.lim, ]

  if(sort.mode == "expected"){

    result <- result[order(result$expected, decreasing = TRUE), ]

  } else if (sort.mode == "observed") {

    result <- result[order(result$observed, decreasing = TRUE), ]

  }

 }

 return(result)

}
```

```
# Script mod.R

mod.error.measures <- function(model, test.set, observed.values, reverse.transformation = NULL){

 predicted <- predict(model, newdata = test.set, type = "response")
```

```
  if(!is.null(reverse.transformation)){

    predicted <- do.call(reverse.transformation, args = list(predicted))

  }

  errors <- observed.values - predicted

  mse <- mean(errors ^ 2)

  rmse <- sqrt(mean(errors ^ 2))

  mae <- mean(abs(errors))

  result <- list(mse = mse, rmse = rmse, mae = mae)

  return(result)

}
```

Tarefas adicionais de pré-processamento

Averiguemos quais são as variáveis com, digamos, mais de 10% de valores em falta:

```
perc <- 0.1

tmp <- apply(carteira, 2, dpp.data.anomalies) # aplica a cada coluna (2) de carteira a função dpp.data.anomalies (dpp.R)
```

Como regra prática, em variáveis com até 10% de valores em falta, podemos descartar as observações envolvidas. Perante percentagens maiores, podemos ter de imputar valores ou *deixar cair* estas variáveis. Por isso, ao ter cerca de 47% de valores em falta, foi decidida a remoção de uma possível variável preditiva em **carteira**, de natureza similar a outras variáveis mantidas.

Coloquemos também de parte quaisquer observações associadas a sinistros causados por catástrofes, uma vez que estes são acontecimentos extremos modelados à parte e que poderão facilmente "inflacionar" fatores multiplicativos que venhamos a calcular no futuro, podendo ainda ser ainda alvo de resseguro:

```
sinistros <- sinistros[sinistros$catastrofe == "", ]
```

Depois deste passo a variável **catastrofe** esgota a sua utilidade e, assim sendo, é preciso removê-la:

```
sinistros <- subset(sinistros, select = -c(catastrofe))
```

Algo de similar acontece em **sinistros**, onde só nos interessam sinistros associados a custos positivos, sendo portanto colocado de parte todo e qualquer sinistro associado a montantes faturados nulos, uma vez que estes são virtualmente indistinguíveis da inexistência de sinistros:

```
sinistros <- sinistros[sinistros$custo_base > 0, ]
```

Dada a natureza deste trabalho e o limite de páginas associado ao texto principal do mesmo, é também recomendada a consulta dos anexos intitulados **Ainda sobre o pré-processamento** e **Determinação da duração da exposição ao risco**.

Por outro lado, e recorrendo à função **anyDuplicated**(), notamos que não há lugar a registos duplicados em nenhum conjunto de dados. Por isso, **não haverá lugar** à remoção de (mais) observações.

Verificada então a qualidade destes conjuntos de dados, é hora de juntá-los num só. Para este efeito, recomenda-se a leitura da secção **Integração de dados** presente nos anexos.

Determinação da duração da exposição ao risco

Variáveis contendo datas também serão úteis para o cálculo da duração de permanência do risco na carteira da ASP. Por isso, vamos optar, tanto em **carteira** como em **sinistros**, por converter os tipos de dados destas variáveis em primeiro lugar, através da função **as.Date**(), para depois a duração pretendida (diretamente útil para o modelo), removendo as variáveis originais no fim.

Por motivos de confidencialidade, é impossível descrever de forma pormenorizada como foi construída a variável do nosso interesse, **duracao_risco_anos**. Porém, isto não impede uma explicação suficientemente detalhada.

Graças às capacidades do R, é possível calcular a duração necessária através de uma mera diferença entre duas datas. Assim:

- A data de início do risco será a de início da apólice, nunca podendo esta data anteceder a data de início de atividade da ASP;
- A data de fim do risco será a de expiração da apólice, nunca podendo esta data ser posterior a 9 de outubro de 2019, dia no qual foram obtidos os conjuntos de dados em análise, sendo ainda esta última data a aplicável em apólices em vigor àquela data.

Note-se que existem indivíduos para os quais o tempo de exposição é muito diminuto, pelo que podemos ter de desconsiderar apólices cuja duração registada da apólice seja muito pequena. O valor de seguida indicado é de 30 dia - ou seja, na verdade, apólices com menos de um mês de vida:

```
dur.min <- 30

carteira <- carteira[carteira$duracao_risco > dur.min, ]
```

Já podemos remover as algumas colunas que nos deixaram de ser úteis, reduzindo a dimensionalidade do conjunto de dados em questão.

Ainda sobre o pré-processamento

Removidas boa parte das variáveis inicialmente ao nosso dispor, podemos focar-nos (mais) agora na conversão de tipos de dados das colunas "sobreviventes", através:

- Da criação de novas variáveis, mais adequadas do que as variáveis já existentes, a partir destas últimas;
- Da recodificação de valores, para uma apresentação mais natural;
- Da remoção e/ou da atribuição[13] de *missing values* a óbvios erros de registo, ou valores que não façam sentido.

Parte do pré-processamento de dados consiste também em diagnosticar a verificação de relações determinísticas que sabemos que existem entre certas variáveis.

De notar que, neste caso e dadas as operações realizadas anteriormente, não precisamos de fazer algo que é, no geral, **necessário** - a conversão de variáveis não devidamente reconhecidas inicialmente para factores ou valores *booleanos*. Por outro lado, notemos que **não há lugar a registos duplicados** em nenhum conjunto de dados, desde o início, pelo que **não haverá lugar** à remoção de (mais) observações em qualquer conjunto de dados.

Integração de dados

Para incorporar em **carteira** a informação contida em **sinistros**, iremos começar por criar um *data.frame* (**freq_com_sinistros**) que faça corresponder aos ID's (**ramo_apol**) de cada sinistrado o número de sinistros para este verificado:

```
tabela <- table(sinistros$ramo_apol)
ID_com_sinistros <- as.numeric(names(tabela))
num_sinistros <- as.numeric(tabela)
freq_com_sinistros <- data.frame(ramo_apol = ID_com_sinistros, num_sinistros = num_sinistros)
```

Se uma dada apólice não for "sinistrada", deve ser adicionada a um outro *data.frame* (**freq_sem_sinistros**) enquanto *falsa sinistrada* com 0 sinistros:

```
I <- carteira$ramo_apol%in%ID_com_sinistros
ID_sem_sinistros <- carteira$ramo_apol[I == FALSE]
freq_sem_sinistros <- data.frame(ramo_apol = ID_sem_sinistros, num_sinistros = 0)
```

Depois disto, devemos proceder à junção destes dois *data.frames* num só:

```
freqs <- rbind(freq_com_sinistros, freq_sem_sinistros)
```

Como iremos modelar separadamente a frequência e a severidade dos sinistros em carteira, fará sentido criar desde já dois conjuntos de dados separados, **frequencias** e **custos** - os quais serão o nosso principal foco a partir de agora:

[13] Sendo que esta atribuição de *missing values* irá naturalmente aumentar a frequência dos mesmos, motivando a possível remoção futura das colunas em questão.

```
frequencias <- merge(carteira, freqs, by = "ramo_apol")
```

Iremos criar ainda um conjunto de dados adicional, **simul**, para simulações futuras:

```
simul <- merge(frequencias, custos, by = "ramo_apol", all.x = TRUE)
simul <- simul[, c("ramo_apol", "num_sinistros", "duracao_risco_anos.x", "custo_base")]
simul$num_sinistros_ano <- simul$num_sinistros / simul$duracao_risco_anos.x
simul <- simul[, c("ramo_apol", "num_sinistros_ano", "custo_base")]

IDs_repetidos <- simul$ramo_apol[duplicated(simul$ramo_apol)] # apólices duplicadas (mais de um sini
stro)
IDs_repetidos <- unique(IDs_repetidos)

for (elem in IDs_repetidos) {

  simul[simul$ramo_apol == elem & !duplicated(simul$ramo_apol), "custo_base"] <- sum(simul$custo_
base[simul$ramo_apol == elem])

}

simul <- simul[!duplicated(simul$ramo_apol), ]

names(simul) <- c("ramo_apol", "num_sinistros_ano", "perdas_totais")

simul$perdas_totais[is.na(simul$perdas_totais)] <- 0

simul$custos_medios_anuais <- simul$perdas_totais / simul$num_sinistros_ano
```

De notar que, pela maneira como foram gerados, os *data.frames* **frequencias** e **custos** estão intimamente ligados a **carteira** e **sinistros**, respetivamente.

Criação de conjuntos de treino, validação e teste

Um passo fundamental, o qual podemos efetivamente considerar parte do pré-processamento de dados, é a realização de separação *train/test* ao nível de observações. Esta separação é extremamente importante, no sentido em que serve para aferir o desempenho dos modelos que possamos construir em *previously unseen data*, isto é, em observações não analisadas na altura da construção de tais modelos.

Nota: A análise exploratória de dados deverá ter apenas como base o conjunto de treino, pois em nenhum momento da construção de modelos devem ser analisadas observações de validação ou de teste.

Serão criadas 3 partições:

- uma de **treino**, compreendendo 80% das observações presentes no conjunto de dados e decomposta em:
    - uma **partição de treino** propriamente dita, com $0.8 \times 0.7 = 0.56$ - 56% das observações;

    - uma **partição de validação** (“teste do treino”) com $0.8 \times 0.3 = 0.24$ - 24% das observações;
- uma de **teste**, compreendendo os restantes 20% das observações presentes no conjunto de dados.

Para este efeito, é fundamental a execução dos seguintes comandos:

```
# Função dpp.train.test.split (dpp.R)

freqs.ind <- dpp.train.test.split(frequencias, 0.8, 0.7)
custos.ind <- dpp.train.test.split(custos, 0.8, 0.7)

frequencias_treino <- frequencias[freqs.ind$train.ind, ]
frequencias_val <- frequencias[freqs.ind$val.ind, ]
frequencias_teste <- frequencias[freqs.ind$test.ind, ]

custos_treino <- custos[custos.ind$train.ind, ]
custos_val <- custos[custos.ind$val.ind, ]
custos_teste <- custos[custos.ind$test.ind, ]
```

De notar que é **absolutamente necessária** a realização das atribuições **car.ind** e **sin.ind**, pois a função **train.test.split** apresenta um *output* **não determinístico** (mesmo tendo sido fixada em cima a *seed*), isto é, apresenta resultados diferentes cada vez que é executada e, assim sendo, a atribuição de um subconjunto (de treino, validação ou teste) a cada observação varia também à medida que tal função é executada. Caso a atribuição em questão não seja realizada, é possível que pelo menos uma observação venha a pertencer a mais do que um destes subconjuntos, o que definitivamente não é o que pretendemos.

Exportação de dados pré-processados

Após a totalidade dos passos acima descritos (de limpeza, integração, redução e transformação de dados), devemos obter conjuntos de dados já pré-processados e prontos para análise. Estes conjuntos podem ser exportados para ficheiros CSV, os quais deverão ser importados novamente em fases posteriores (exploração, modelação); assim, mantemos tais conjuntos imutáveis (depois de criados).

Assim sendo, iremos escrever os *data.frames* gerados nos diretórios relevantes, nomeadamente **./pipeline/2_aed/1_entrada** e **./pipeline/3_modelos/1_entrada**. Tal missão pode ser cumprida através de:

```
### Para frequências

# A análise exploratória de dados terá apenas como base o conjunto de treino, pois em nenhum momento da construção de modelos observações de validação ou de teste devem ser analisadas

write.csv2(frequencias_treino, "./pipeline/2_aed/1_entrada/frequencias_treino.csv")

# Tendo criado este ficheiro, não faz sentido criá-lo de novo, por isso, vamos apenas copiá-lo

file.copy("./pipeline/2_aed/1_entrada/frequencias_treino.csv", "./pipeline/3_modelos/1_entrada/frequencia
```

```
s_treino.csv", overwrite = fazer.split)

# Na fase de avaliação de modelos faremos também uso dos conjuntos de validação e de teste

write.csv2(frequencias_val, "./pipeline/3_modelos/1_entrada/frequencias_val.csv")
write.csv2(frequencias_teste, "./pipeline/3_modelos/1_entrada/frequencias_teste.csv")

### Para custos

write.csv2(custos_treino, "./pipeline/2_aed/1_entrada/custos_treino.csv")

file.copy("./pipeline/2_aed/1_entrada/custos_treino.csv", "./pipeline/3_modelos/1_entrada/custos_treino.csv", overwrite = fazer.split)

write.csv2(custos_val, "./pipeline/3_modelos/1_entrada/custos_val.csv")
write.csv2(custos_teste, "./pipeline/3_modelos/1_entrada/custos_teste.csv")

write.csv2(simul, "./pipeline/3_modelos/1_entrada/simul.csv")
```

Análise exploratória de dados

Tendo sido completada a etapa de pré-processamento de dados, passemos à exploração dos mesmos, importando-os através de:

```
frequencias_treino <- read.csv2("./pipeline/2_aed/1_entrada/frequencias_treino.csv")

custos_treino <- read.csv2("./pipeline/2_aed/1_entrada/custos_treino.csv")
```

Por alguma razão foi incluída, no processo de exportação e posterior importação de dados (tanto em frequencias como em custos), uma coluna **X** que **não é do nosso interesse**, pelo que iremos removê-la:

```
frequencias_treino <- subset(frequencias_treino, select = -c(X))

custos_treino <- subset(custos_treino, select = -c(X))
```

Como é injusto comparar a sinistralidade de apólices com durações distintas, iremos criar em **frequencias_treino** uma nova variável que melhor reflita esta realidade:

```
frequencias_treino$num_sinistros_ano <- frequencias_treino$num_sinistros/frequencias_treino$duracao_risco_anos
```

Como a variável **custo_base** poderá no futuro ser alvo de uma transformação logarítmica (no momento de modelação), podemos também criar desde já uma nova variável:

```
custos_treino$custo_base_log <- log(custos_treino$custo_base)
```

Análises podem ser **univariadas** ou **multivariadas**, consoante envolvem uma ou mais variáveis aleatórias (respetivamente) sendo certo que, em qualquer dos casos, os tipos de explorações a efetuar irão depender do tipo de dados das variáveis em questão.

Tendo em mente o pré-processamento já realizado, e passando à análise dos dados, podemos obter de estatísticas amostrais univariadas através do comando **summary()**. No entanto, o *output* deste comando talvez não seja tão completo quanto seria desejável, pelo menos para variáveis numéricas. Por isso, há valor no recurso às funções presentes no *script* **eda.R**, nomeadamente as funções **eda.univariate.numeric.stats**(), **eda.conditional.categorical.barplots**(), **eda.conditional.numeric.groupby.plots**() , **eda.conditional.numeric.groupby.stats()**, e **eda.joint.numeric.plots()**.

Analisando a distribuição das frequências de sinistros (com 123449 observações), podemos notar:

- Que o cenário mais comum é o de **não-ocorrência de sinistros**;
- Que **não parece haver sobredispersão**, pois a média amostral está próxima da variância;
- Que esta distribuição é assimétrica, possuindo uma cauda à direita.

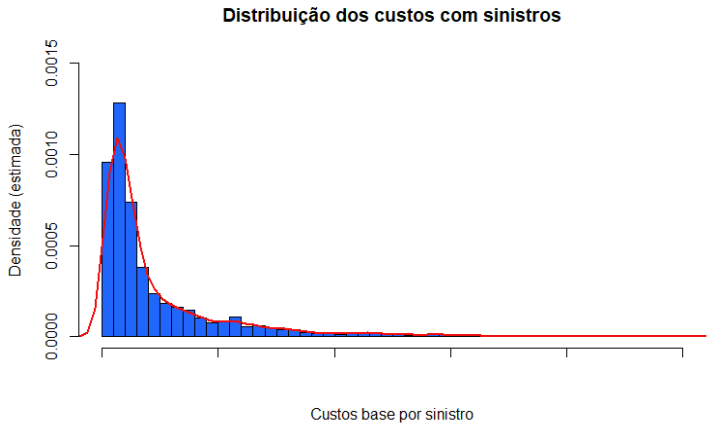

Figura A.3 – Distribuição dos custos base com sinistros

Notamos, na distribuição empírica dos custos base (2390 observações), que a **dispersão é muito elevada** (pois o desvio-padrão é mais do dobro da média) e a **distribuição muito assimétrica** (pois o coeficiente de assimetria amostral é de cerca de 13.63, a média excede o dobro da mediana, e apenas cerca de um quarto dos custos observados superam o custo médio).

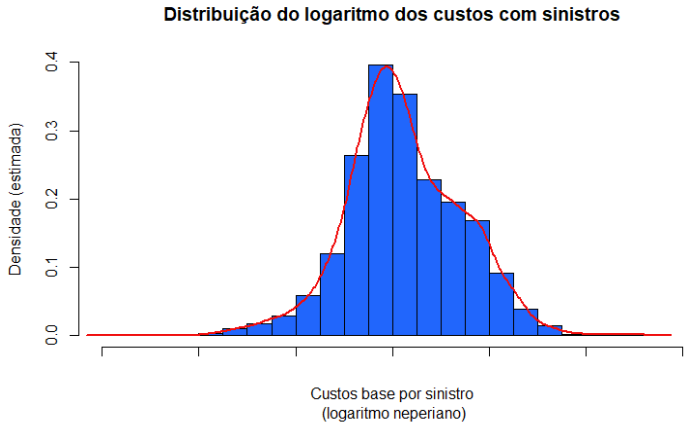

Figura A.4 – Distribuição do logaritmo neperiano dos custos base com sinistros

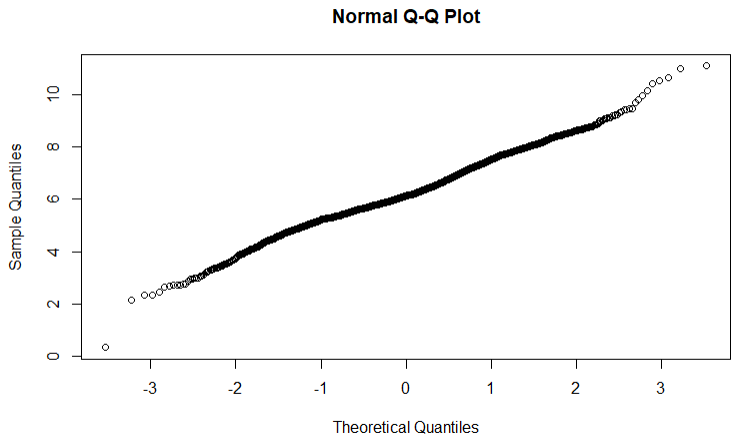

Figura A.5 – Q-Q plot: gráfico que compara os quantis teoricamente aplicáveis da distribuição Normal com os quantis da distribuição empírica do logaritmo neperiano dos custos com sinistros

Notamos ainda que **o logaritmo neperiano dos custos base com sinistros parece seguir aproximadamente uma distribuição Normal** - a densidade estimada possui o formato de *curva de sino*, os pontos parecem estar dispostos ao longo de uma reta no *QQ-plot*, e o coeficiente de achatamento/curtose observado (de cerca de 3.615) parece estar próximo de 3.

Ao nível das **variáveis qualitativas** ou **categóricas**, não há muitos resultados de interesse a explorar. No entanto, e ao nível da **variável 1**, notamos que:

- Tanto a maioria dos imóveis protegidos como a maioria dos imóveis envolvidos em sinistros se encontram nas categorias **A** (2.º par de barras), **M** (1.º par), **Q** (5.º par) e **D** (3.º par);
- Parece ainda haver uma **maior** propensão para sinistros nas categorias **M** e **D**, e uma **menor** propensão nas categorias **A** e **Q**.

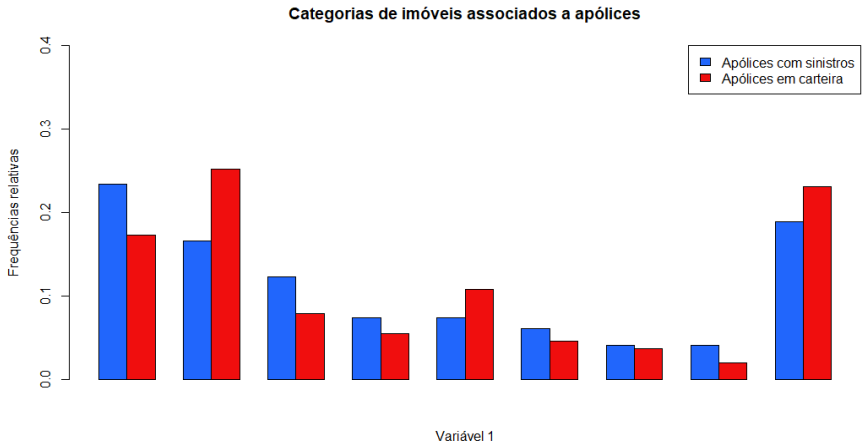

Figura A.6 – Número de apólices (com sinistros *vs* no todo) por categoria da variável 1

Adicionalmente, observamos que duas das variáveis ao nosso dispor apresentam apenas valores associados a uma classe, pelo que na verdade não variam, podendo por isso ser descartadas, visto não nos ser útil para fins de exploração ou de modelação.

Passemos àquela que será, muito provavelmente, a parte mais importante desta fase. Falamos das **análises multivariadas**, sobretudo **bivariadas**, nas quais iremos considerar relações entre:

1. As **variáveis a prever** ou explicar - **num_sinistros_ano** e **custo_base**;
2. As **variáveis preditivas** ou explicativas, **var1**, **var2**, **var3**, **var4, var5**, **var6 e var7**.

Comecemos por estudar relações que envolvam a variável **var3**.

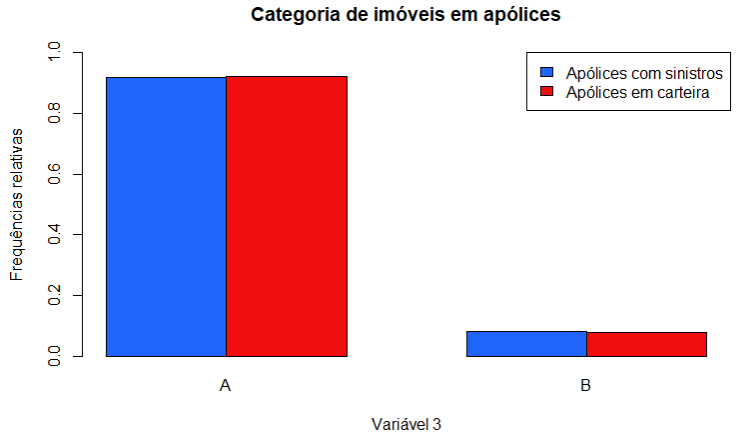

Figura A.7 – Número de apólices (com sinistros *vs* no todo) por categoria da variável 3

Verificamos que as proporções de apólices em carteira nas categorias A e B se mantêm no conjunto de apólices com sinistros. Por outras palavras, as diferenças na **frequência** de sinistros nestas duas categorias é (quase) invisível, pelo que talvez esta variável não nos seja útil na modelação de frequências.

Já ao nível de **custos** base médios com sinistros, verificamos diferenças mais significativas entre estas categorias, sendo estes **maiores na categoria B**, e **menores na categoria A**. Apesar disto, e ainda ao nível de custos, parecem também haver bem mais *outliers* na categoria A.

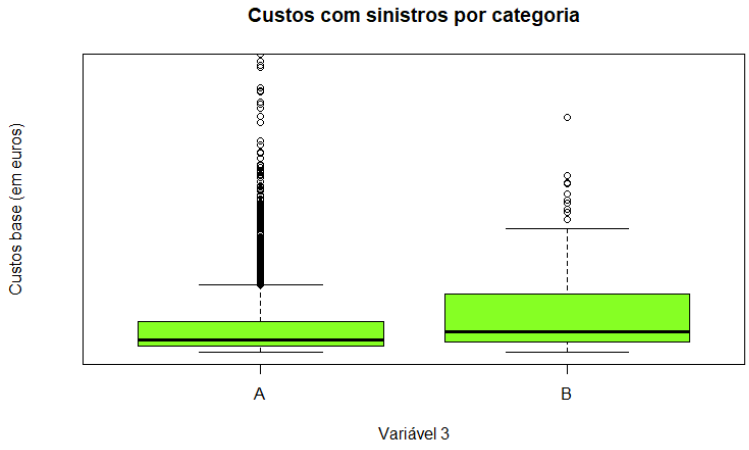

Figura A.8 – Custos com sinistros por categoria da variável 3

Passando para a variável **var6**, e ao nível da frequência de sinistros, temos:

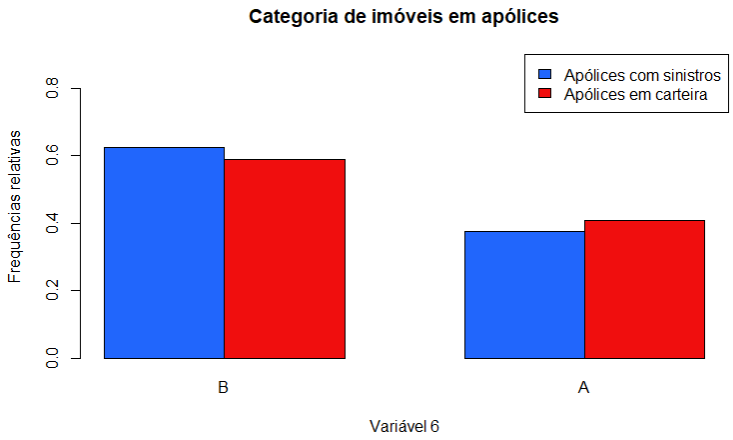

Figura A.9 – Número de apólices (com sinistros *vs* no todo) por categoria da variável 6

Podemos notar que a frequência de sinistros é ligeiramente maior em apólices na categoria **B** desta variável, sendo ainda os custos com sinistros significativamente maiores na mesma:

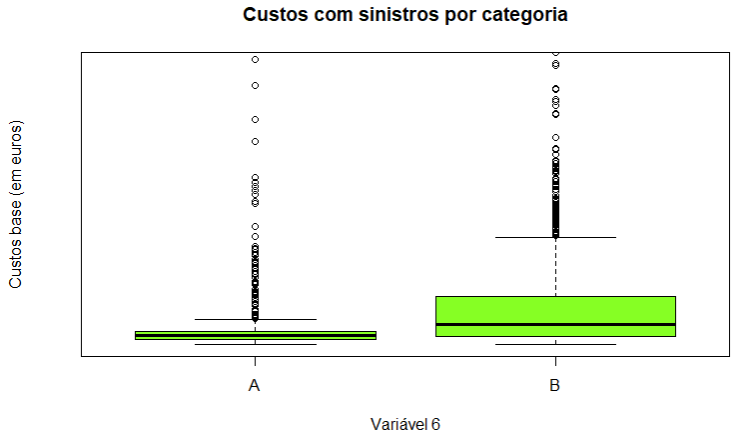

Figura A.10 - Custos com sinistros por categoria da variável 6

Já em **var7**, e ao nível de frequências:

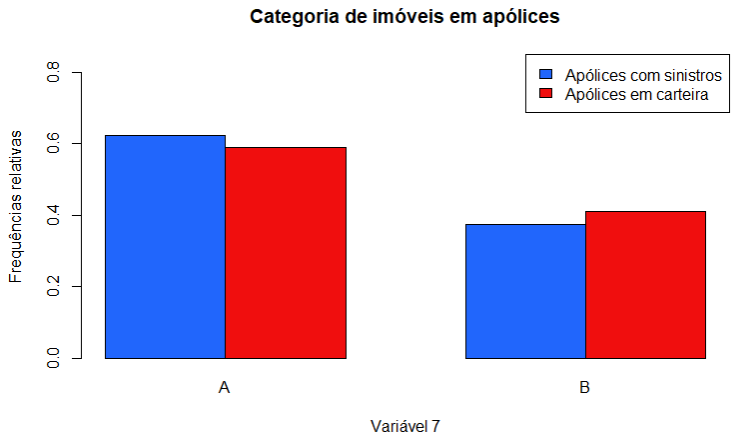

Figura A.11 - Número de apólices (com sinistros *vs* no todo) por categoria da variável 7

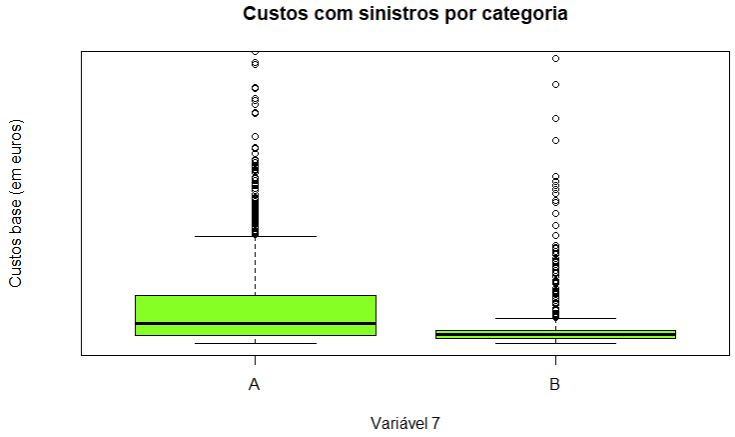

Figura A.12 - Custos com sinistros por categoria da variável 7

A última variável qualitativa do nosso interesse para estudos bivariados é **var1**.

Tabela A.7 - Número de sinistros por apólice, por ano e por categoria da variável 1

Categoria	Percentagem de apólices	Média por categoria	Variância por categoria
R	1.9595 %	0.031207	0.026396
D	7.8462 %	0.024780	0.024214
J	0.7517 %	0.024778	0.020433
S	1.4233 %	0.023390	0.020028
T	2.3726 %	0.022149	0.020324
M	17.2436 %	0.021910	0.025495

B	5.4808 %	0.020945	0.019989
F	1.6719 %	0.020279	0.018028
E	1.0466 %	0.020125	0.017380
G	4.6141 %	0.020113	0.018129
K	3.6801 %	0.017477	0.016584
L	0.8457 %	0.015973	0.014837
P	3.7206 %	0.012383	0.010180
N	0.7461 %	0.010818	0.010299
Q	10.8101 %	0.010510	0.009580
A	25.2226 %	0.010217	0.012696
O	2.1750 %	0.008498	0.007536
C	1.0101 %	0.005669	0.004424
I	6.0689 %	0.004976	0.004097
H	1.3107 %	0.004974	0.004002

Daqui concluimos que algumas categorias são **mais propensas** à ocorrência de sinistros, nomeadamente as categorias **R, D, J** e **S**, sendo outras **menos propensas**, como é o caso das categorias **Q, A, O, C, I** e **H**. Já ao nível de custos, temos:

Tabela A.8 - Custo base por sinistro e por categoria da variável 1

Categoria	Percentagem de sinistros	Média por categoria	Desvio-padrão por categoria
N	0.6276 %	1924.0487	3208.7833
O	1.0879 %	1622.1746	2520.5053
A	16.4435 %	1491.5283	3879.8468
Q	6.9456 %	1394.8039	2412.0819
D	12.5523 %	1389.9579	3933.7600
I	2.5105 %	1216.6327	1720.7679
P	3.0126 %	1185.8029	2353.8902
B	7.4059 %	997.2598	4494.3840
M	24.3096 %	969.0043	1404.7882

K	3.8912 %	949.9548	1667.3495
L	0.5858 %	925.7807	1176.0204
H	0.5439 %	924.2269	1043.1761
S	1.7992 %	913.9516	1483.7233
E	1.0460 %	830.6804	1317.4963
G	5.3556 %	812.7612	1040.2749
R	4.1841 %	796.2873	991.2120
F	2.6360 %	725.5849	790.9599
C	0.4184 %	702.3390	886.8079
J	1.1297 %	672.0752	762.9624
T	3.5146 %	651.2154	1060.1805

Os valores **mais elevados** são encontrados nas categorias **N**, **O** e **A**. Por outro lado, a severidade dos sinistros tende a ser **menor** nas categorias **T, J e C**.

Passando para variáveis quantitativas como **var5**, verificamos que a relação existente entre esta variável e as variáveis a prever não é muito forte:

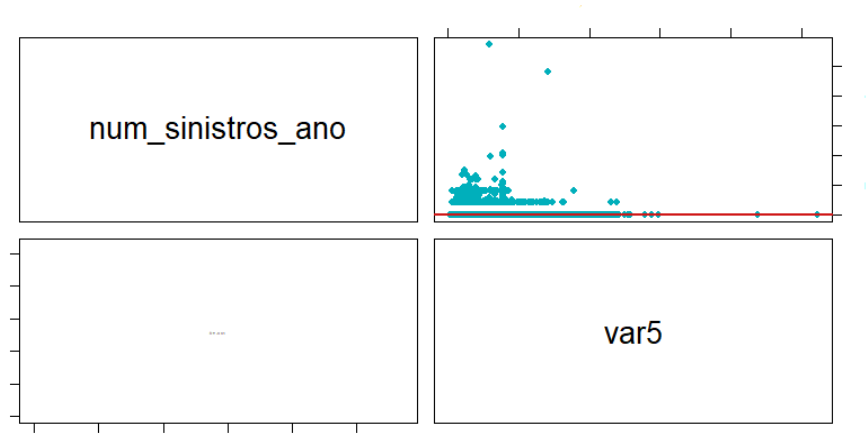

Figura A.13 – Relação existente entre o número de sinistros por apólice e por ano e a variável 5

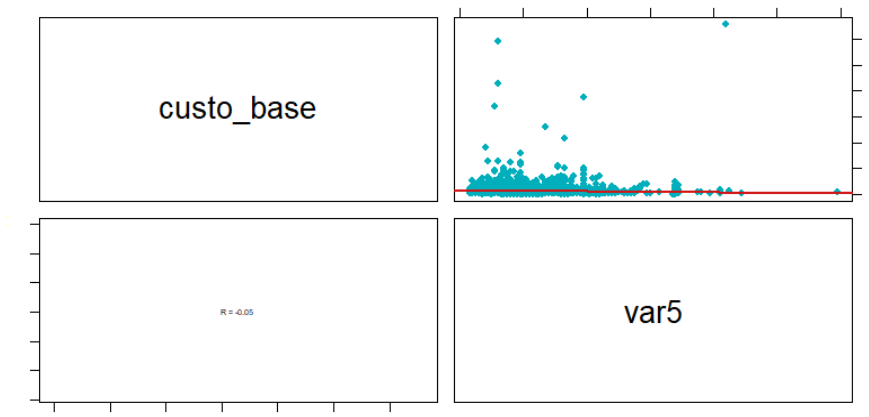

Figura A.14 – Relação existente entre o custo base por sinistro e a variável 5

Já em relação a **var2**, não parece haver nenhum destaque minimamente significativo, uma vez que:

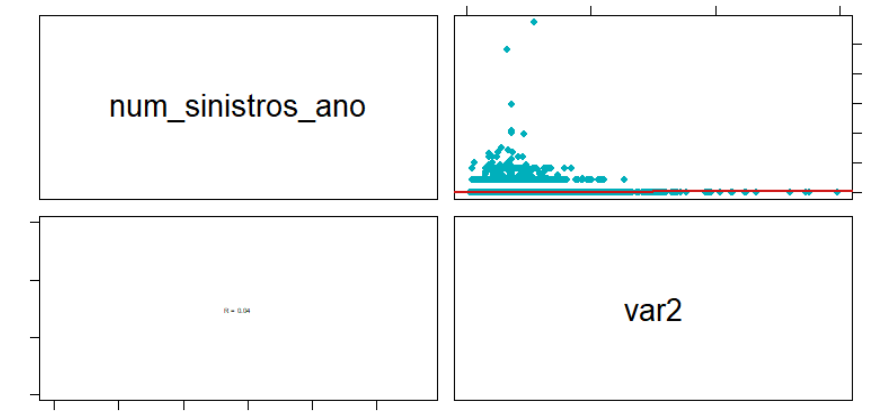

Figura A.15 – Relação existente entre o número de sinistros por apólice e por ano e a variável 2

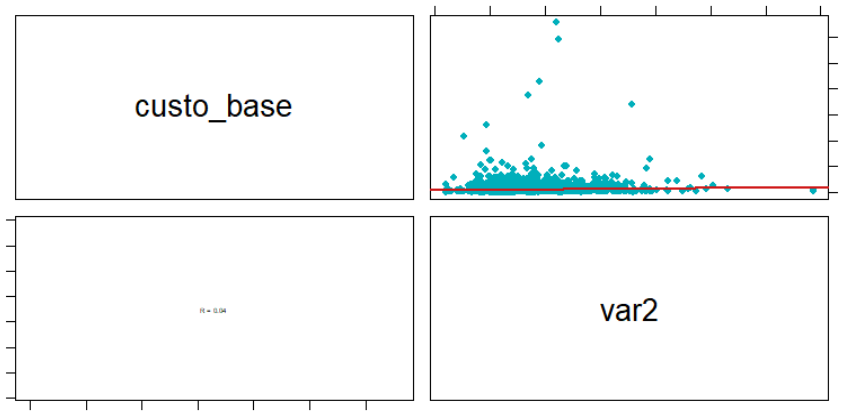

Figura A.16 – Relação existente entre o custo base por sinistro e a variável 2

Por último, e ao nível de **var4**, temos uma relação algo interessante entre esta variável e **custo_base**, algo que já não se verifica em **num_sinistros_ano**:

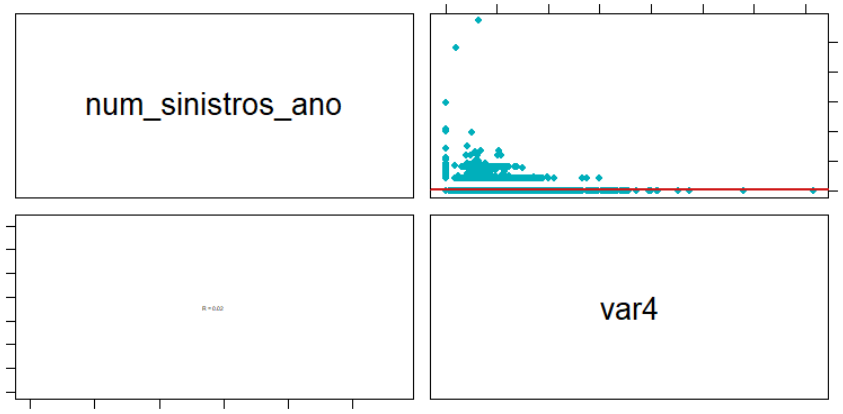

Figura A.17 – Relação existente entre o número de sinistros por apólice e por ano e a variável 4

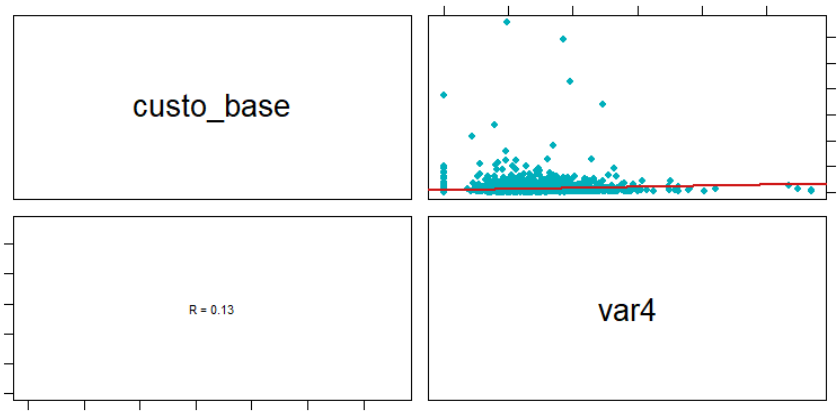

Figura A.18 – Relação existente entre o custo base por sinistro e a variável 4

Observamos que parece existir uma relação crescente entre **var4** e **custo_base**, a qual faz sentido, economicamente falando.

Podemos concluir, através destas análises exploratórias:

- Que de entre as variáveis que podem explicar tanto a frequência como a sinistralidade, notamos algumas relações ténues (por exemplo, entre uma destas variáveis e a **variável 2**);
- Que outras relações são definitivamente de maior interesse - por exemplo, entre pelo menos uma destas variáveis e as variáveis 1, 3, 4, 5, 6 e 7.

Modelação – alguns passos prévios

Iremos procurar categorizar as variáveis quantitativas contínuas ao nosso dispor, de maneira a que cada intervalo ou classe de valores originais tenha aproximadamente o mesmo número de observações ou esteja de outra forma suficientemente representado, para evitar a construção de classes desequilibradas. Daí o recurso à função **quantile()**.

A categorização de **var2** teve início considerando, para os limites de cada classe, os seus quartis (em adição ao máximo e ao mínimo). Verificou-se que o 1º. e o 4º quartil de **var2** poderiam, também eles, ser separados ou divididos em dois, para captar o comportamento de apólices com valores extremos para esta variável, até porque o ajuste de modelos de forma iterativa revelava que, de facto, havia espaço para a construção de mais categorias neste fator, dado que tais novas categorias eram estatisticamente significativas.

Podemos então considerar, para **var2**, uma categorização com os seguintes resultados:

Tabela A.9 – Categorização da variável 2

Escalão	Percentagem de apólices	Percentagem de sinistros

A	3.880 %	1.339 %
B	16.323 %	10.753 %
C	22.377 %	21.590 %
D	44.568 %	47.573 %
E	10.429 %	13.849 %
F	2.423 %	4.895 %

Os diversos fatores tarifários considerados para a modelação foram, portanto, construídos com base na partição de variáveis contínuas da forma mais equitativa, para evitar que qualquer célula tarifária (resultante da interseção de níveis dos diversos fatores tarifários) esteja, à partida, pouco representada, isto é, esteja associada a uma contagem baixa. Podem haver, ainda assim, células tarifárias pouco representadas, mas ao menos estamos a dificultar a obtenção de tal desfecho.

É do nosso interesse a escolha de bons níveis base para cada fator tarifário, uma vez que quanto maior a frequência dos mesmos, mais sólida será a nossa análise estatística. Por isso, iremos, em determinadas variáveis, agrupar categorias ou alterar categorias base.

Todas as alterações realizadas neste momento no conjunto de treino terão de ser também aplicadas posteriormente nos conjuntos de validação e de teste.

Testes à distribuição das frequências de sinistros

```
dados <- frequencias_treino$num_sinistros[frequencias_treino$var1 == "A" & frequencias_treino$var2 = = "D" & frequencias_treino$var3 == "A" & frequencias_treino$var5 == "C" & frequencias_treino$var6 = = "B" & frequencias_treino$var7 == "A"]

media.freqs <- mean(dados)

freqs.obs.sinistros <- table(dados)
freqs.esp.poisson <- as.table(dpois(0:2, lambda = media.freqs))

row.names(freqs.esp.poisson) <- 0:2

freqs.obs.sinistros[2] <- freqs.obs.sinistros[2] + freqs.obs.sinistros[3]
freqs.obs.sinistros <- freqs.obs.sinistros[1:2]
freqs.esp.poisson[2] <- ppois(0, media.freqs, F)
freqs.esp.poisson <- freqs.esp.poisson[1:2]

freqs.esp.poisson <- freqs.esp.poisson * length(dados)

freqs.obs.sinistros <- as.data.frame(freqs.obs.sinistros)
freqs.esp.poisson <- as.data.frame(freqs.esp.poisson)
```

```
freqs <- merge(freqs.obs.sinistros, freqs.esp.poisson, by.x = "dados", by.y = "Var1", all = TRUE)

colnames(freqs) <- c("", "freqs.obs.sinistros", "freqs.esp.poisson")

freqs[is.na(freqs)] <- 0

freqs <- as.matrix(freqs)

freqs <- freqs[, -1]

freqs <- apply(freqs, 2, as.numeric)

freqs <- cbind(freqs, (freqs[, "freqs.obs.sinistros"] - freqs[, "freqs.esp.poisson"])^2 / freqs[, "freqs.esp.poisson"])
colnames(freqs) <- c("freqs.obs.sinistros", "freqs.esp.poisson", "estat.teste")
freqs
```

```
##      freqs.obs.sinistros freqs.esp.poisson estat.teste
## [1,]                 374         373.06408 0.002347978
## [2,]                   6           6.93592 0.126291285
```

```
pchisq(sum(freqs[, "estat.teste"]), df = 4-1-1, lower.tail = FALSE)
```

```
## [1] 0.9377052
```

Testes à distribuição dos resíduos da regressão linear

```
classes <- qnorm(seq(0, 1, 0.1), 0, sd(modelo.custos.lognormal.2$residuals))
table(cut(modelo.custos.lognormal.2$residuals, classes))
```

```
##
##   (-Inf,-1.47] (-1.47,-0.962]  (-0.962,-0.6]   (-0.6,-0.29]     (-0.29,0]
##            196            214            236            258            289
##      (0,0.29]     (0.29,0.6]    (0.6,0.962]   (0.962,1.47]    (1.47, Inf]
##            256            284            218            224            215
```

```
stat <- chisq.test(table(cut(modelo.custos.lognormal.2$residuals, classes)), p = rep(0.1, 10))$statistic

pchisq(stat, df = 10 - 2 - 1, lower.tail = FALSE)
```

```
##    X-squared
## 4.227062e-06
```

Princípios de cálculo de prémios. Definição de *loadings*

Um princípio de cálculo de prémio mais não é do que uma regra que permite calcular o prémio puro ou prémio de risco associado a uma apólice. Sejam consideradas as seguintes variáveis aleatórias:

- S - Valor total das indemnizações a serem pagas pela seguradora, numa dada apólice e num dado período;
- N - Número total de sinistros numa apólice (escolhida ao acaso);
- X - Valor de uma indemnização individual (escolhido ao acaso);

bem como os seguintes valores:

- $\Pi(X)$ - Prémio puro (com uma possível margem de segurança embutida) associado ao risco X;
- α (ou λ ou θ) - Carga de segurança.

Vamos ainda assumir que a variável aleatória S envolvida possui valor esperado $\mathbb{E}(S)$ e variância $\mathrm{var}(S)$. Se N e X forem variáveis aleatórias independentes, ou pelo menos não-correlacionadas, teremos

$$\mathbb{E}(S) = \mathbb{E}(N)\mathbb{E}(X)$$

O primeiro princípio de cálculo de prémios que veremos será o princípio do prémio puro, de natureza mais simples, segundo o qual o prémio a pagar corresponde ao valor justo ou valor atuarial do risco, isto é,

$$\Pi(S) = \mathbb{E}(S)$$

A aplicação deste prémio sem margens de segurança levará a insuficiências, ou seja, à ruína da seguradora, independentemente de quão elevadas sejam as suas reservas iniciais (finitas). Adicionalmente, o prémio estimado com base em dados passados pode não refletir adequadamente custos com indemnizações futuras.

Este método é muitas vezes usado, através da aplicação de *loadings*, isto é, de margens de segurança que visam mitigar as limitações do mesmo. Assim se obtém o próximo princípio, o princípio do valor esperado com margem de segurança, de acordo com o qual

$$\Pi(S) = (1+\alpha)\mathbb{E}(S), \alpha \geq 0$$

Neste princípio, α corresponde à margem de segurança relativa, e $\alpha\mathbb{E}(X)$ corresponde à margem de segurança total. Este é o princípio mais utilizado na prática, o que não faz com que seja universalmente melhor do que os restantes, até porque não tem em conta o grau das possíveis flutuações de X, o que motiva o surgimento de outros princípios.

Por sua vez, os princípios da variância e do desvio-padrão visam ter em conta não só o valor esperado das perdas, mas também o grau de dispersão das mesmas, sendo respetivamente dados por:

$$\Pi(S) = \mathbb{E}(S) + \alpha var(S), \alpha \geq 0$$

$$\Pi(S) = \mathbb{E}(S) + \alpha\sqrt{var(S)}, \alpha \geq 0$$

Estes dois princípios visam cobrir perdas com elevada probabilidade e, de maneira intencional, geram prémios que excedem as perdas esperadas, servido a diferença entre estes dois valores de margem de segurança, isto é, de almofada financeira que servirá para fazer face a experiências adversas. A única exceção ocorre quando $\alpha = 0$, cenário que resulta no princípio do prémio puro.

A preocupação com a dispersão prende-se com a possível ocorrência de perdas extremamente elevadas e, neste contexto, o princípio do quantil é também interessante; de acordo com este princípio, se S for uma variável aleatória absolutamente contínua (o que tende a ser o caso), ter-se-á

$$\Pi(S) = F_S^{-1}(1-\varepsilon), \varepsilon \in [0,1]$$

O prémio calculado de acordo com este princípio é então o quantil de ordem $(1-\varepsilon)$ da distribuição de perdas em estudo, ou seja, é o valor que se espera ser suficiente para cobrir as perdas observadas em $(1-\varepsilon) \times 100\%$ dos anos (ou outros períodos considerados), desde que esta distribuição se mantenha inalterada no tempo. Valores razoáveis para ε tendem a situar-se entre 1% e 5%.